

Metacognizione: il modello del modello

Terza monografia del ciclo *Architettura della mente*

Nota sul ciclo

Questa monografia è la terza di quattro che condividono un impianto tecnico comune pur adottando apparati formali distinti, ciascuno scelto per la sua affinità strutturale con l'oggetto di studio.

La prima — *Il malinteso come struttura cognitiva* — ha trattato il malinteso non come fallimento della comunicazione ma come il meccanismo fondamentale attraverso cui sistemi con rappresentazioni diverse producono e negoziano significato. Il suo apparato è venuto dalla teoria dei giochi della comunicazione e dalla semantica formale: lo spazio degli stati mentali come varietà su cui gli agenti mantengono distribuzioni di credenza, l'equilibrio di Nash come struttura del significato condiviso, il malinteso come punto fuori equilibrio in cui le distribuzioni degli agenti hanno divergenza di Kullback-Leibler positiva.

La seconda — *Pensare senza parole* — ha costruito il modello del substrato subpropositivo come sistema di rappresentazione distribuita in uno spazio di attivazione $\mathbb{R}^n$ di cui il linguaggio è un codec lossy con perdita informazionale misurabile. Il sistema dinamico accoppiato substrato-linguaggio ha mostrato che Φ^{-1} è un operatore che rimodella la geometria di $\mathbb{R}^n$ attraverso l'uso ripetuto, producendo la co-evoluzione il cui attrattore determina la struttura cognitiva stabile dell'agente esperto.

La terza monografia — il testo che segue — affronta la metacognizione come problema di compressione di secondo ordine: il sistema mantiene un modello compresso dei propri modelli compressi, e la qualità di quel meta-modello determina l'allocazione delle risorse cognitive. L'apparato formale viene dalla teoria bayesiana gerarchica, dai teoremi di impossibilità per l'auto-conoscenza formale — Gödel, Tarski, Löb, Rice, Chaitin, Lawvere nel suo schema categoriale unificante — e dai teoremi di impossibilità per l'aggregazione delle preferenze di Arrow e Sen come istanze della stessa struttura nell'auto-modellazione dei sistemi multi-agente.

La quarta — *L'attenzione come operatore di selezione del gradiente* — chiuderà il ciclo trattando l'attenzione non come risorsa scarsa ma come selettore di quale prior viene

aggiornato, con le conseguenze che questo ha per l'economia politica formale dell'evoluzione collettiva delle credenze.

Le quattro monografie formano un ciclo coerente: il malinteso è il fallimento del meta-modello dell'altro; il substrato subpropositivo è il dominio che il meta-modello non può comprimere senza distorsione massima; la metacognizione è il meccanismo che determina quanto di quel dominio viene allocato all'elaborazione esplicita; l'attenzione è il selettore che, operando sul gradiente dell'aggiornamento bayesiano, orienta l'intera architettura verso una rappresentazione del mondo che porta l'impronta di chi la controlla.

Proemio: Il problema dell'auto-conoscenza

C'è un momento preciso in cui l'epistemologia cessa di essere la teoria della conoscenza del mondo e diventa la teoria della conoscenza della conoscenza. Quel momento è antico — Aristotele distingue il pensiero che pensa se stesso nel libro Lambda della Metafisica, Kant dedica la Critica della ragion pura alla capacità della ragione di conoscere i propri limiti, Husserl fonda l'intera fenomenologia sull'intenzionalità riflessiva — ma la sua formalizzazione rigorosa è novecentesca, e le sue conseguenze per la scienza cognitiva sono ancora, in larga parte, da trarre.

Il punto di partenza è un risultato che la logica matematica ha stabilito con certezza e che la filosofia della mente ha troppo a lungo trattato come un'anomalia tecnica anziché come un vincolo strutturale fondamentale: un sistema sufficientemente potente da modellare se stesso è soggetto a teoremi di impossibilità che limitano in modo preciso e irriducibile la fedeltà della propria auto-rappresentazione. Il risultato di Gödel — che ogni sistema formale coerente e sufficientemente espressivo contiene enunciati veri ma indimostrabili al proprio interno — è il più noto di questi teoremi, ma è anche il più frainteso. Trattato come un'oscura curiosità logica, viene spesso invocato retoricamente (la coscienza supera il meccanismo, l'intuizione supera il calcolo) senza che ne venga estratto il contenuto formale preciso. Il contenuto preciso è che esiste un livello irriducibile di opacità del sistema rispetto a se stesso: una distanza positiva e non annullabile tra la struttura del sistema e la rappresentazione che il sistema ha di quella struttura.

Questo limite non è un difetto riparabile con maggiore potenza computazionale o maggiore accuratezza introspettiva: è un limite strutturale che vale per qualsiasi sistema con le caratteristiche richieste, indipendentemente dalla sua sofisticazione. La metacognizione — la capacità del sistema cognitivo di modellare se stesso — opera necessariamente su una rappresentazione compressa e distorta del proprio oggetto. La domanda scientificamente rilevante non è se questo limite esiste — esiste, ed è dimostrabile — ma quale struttura ha la distorsione che introduce, come si misura, e come un sistema razionale dovrebbe comportarsi alla luce della propria auto-opacità.

La risposta a queste domande richiede tre ingredienti che le trattazioni precedenti della metacognizione non hanno combinato sistematicamente: un apparato formale che catturi la struttura gerarchica della rappresentazione di secondo ordine (la teoria bayesiana gerarchica); una precisa catalogazione dei teoremi di impossibilità che limitano l'auto-conoscenza formale (da Gödel a Lawvere, con l'inclusione di Rice, Chaitin, Löb e i risultati di Wolpert-Macready sul no-free-lunch); e un'estensione dell'analisi ai sistemi multi-agente, dove i teoremi di impossibilità per l'aggregazione delle preferenze (Arrow, Sen) emergono come istanze della medesima struttura formale nel dominio dell'auto-modellazione collettiva.

L'argomento si sviluppa in cinque movimenti. Il primo costruisce il framework bayesiano gerarchico per la metacognizione e ne identifica la struttura come limite proiettivo di una torre di distribuzioni di credenza. Il secondo analizza i teoremi di impossibilità per l'auto-conoscenza formale, unificandoli attraverso il teorema di punto fisso di Lawvere come schema categoriale comune. Il terzo affronta la metacognizione come problema di compressione di secondo ordine e ne costruisce la funzione di tasso-distorsione. Il quarto estende l'analisi all'allocazione delle risorse cognitive come problema di ottimizzazione vincolata sotto auto-opacità. Il quinto porta il framework ai sistemi multi-agente, connettendo i risultati di impossibilità per l'auto-modellazione individuale con quelli per l'aggregazione collettiva.

Capitolo I — La struttura bayesiana gerarchica della metacognizione

I.1 Dalla cognizione alla metacognizione: il salto di livello

Nella seconda monografia del ciclo, la cognizione è stata formalizzata come la traiettoria di un vettore $v(t)$ nello spazio di attivazione $\mathbb{R}^n$, guidata dall'equazione dinamica accoppiata tra substrato e linguaggio. Questo framework tratta il sistema cognitivo come un processo nel suo spazio di stati — una curva nella varietà cognitiva $M \subset \mathbb{R}^n$. La metacognizione richiede un ulteriore passaggio: il sistema deve non solo percorrere M ma costruire e mantenere una rappresentazione di M , della propria posizione su M , e della propria dinamica su M . Non si tratta di muoversi in uno spazio ma di avere una mappa di quello spazio — e, crucialmente, una mappa della qualità della propria mappa.

La formalizzazione naturale di questa struttura è bayesiana. Il sistema cognitivo non ha accesso diretto alla propria struttura: osserva le proprie uscite e i propri stati accessibili alla coscienza, e da queste osservazioni infere la propria struttura sottostante. In termini formali, il sistema ha una distribuzione di credenza sul proprio spazio degli stati cognitivi — non un singolo punto in $\mathbb{R}^n$, ma una distribuzione di probabilità su $\mathbb{R}^n$ che rappresenta la stima del sistema circa la propria posizione nell'architettura cognitiva.

Sia Θ lo spazio dei parametri che descrive l'architettura cognitiva del sistema: i pesi sinaptici W del substrato, i parametri della competenza linguistica m_lang , le soglie e le connessioni dei sistemi di attenzione e controllo. Lo stato metacognitivo del sistema al tempo t è:

$$\pi_t(\theta) = P(\Theta = \theta \mid O_{\{0:t\}})$$

la distribuzione di probabilità su Θ condizionata all'insieme $O_{\{0:t\}}$ delle osservazioni accessibili al sistema fino al tempo t . Questa distribuzione — il *prior metacognitivo* del sistema — è la rappresentazione che il sistema ha di se stesso.

I.2 La torre bayesiana gerarchica e il suo limite proiettivo

La struttura bayesiana si rivela immediatamente gerarchica. Il sistema ha un prior sui propri parametri cognitivi $\pi_t(\theta)$: questo è il primo livello della gerarchia, il meta-modello. Ma il prior π_t è esso stesso un oggetto che il sistema può modellare: si può considerare la distribuzione del sistema sulle proprie distribuzioni metacognitive — la distribuzione di secondo ordine sulla bontà del proprio meta-modello. E questo secondo livello induce un terzo, e così via.

Formalmente, si definisce la torre bayesiana gerarchica come la sequenza:

$$\begin{aligned} \pi^0 &= \text{distribuzione su } \Theta \text{ (modello dei parametri cognitivi)} \\ \pi^1 &= \text{distribuzione su } \Pi^0 \text{ (meta-modello: distribuzione sulle distribuzioni su } \Theta) \\ \pi^2 &= \text{distribuzione su } \Pi^1 \text{ (meta-meta-modello)} \\ &\dots \\ \pi^k &= \text{distribuzione su } \Pi^{k-1} \end{aligned}$$

dove Π^k è lo spazio delle distribuzioni di probabilità su Π^{k-1} . La torre è una sequenza di spazi di funzioni — lo spazio dei funzionali sul livello sottostante — con morfismi di marginalizzazione che collegano i livelli:

$$\text{marg}_{\{k,k-1\}} : \Pi^k \rightarrow \Pi^{k-1}, \quad \text{marg}_{\{k,k-1\}}(\pi^k)(A) = \int \pi^{k-1}(A) d\pi^k(\pi^{k-1})$$

La marginalizzazione discende la torre: da un prior di livello k si ricava il prior marginale di livello $k-1$ integrando su tutti i possibili valori del livello superiore.

La struttura matematica della torre bayesiana gerarchica è quella di un sistema proiettivo nella categoria delle misure di probabilità. Si definisce il *limite proiettivo* della torre come lo spazio:

$$\Pi^\infty = \lim_{\leftarrow} \{\leftarrow\} (\Pi^0 \leftarrow \Pi^1 \leftarrow \Pi^2 \leftarrow \dots)$$

Il teorema di Kolmogorov assicura, sotto condizioni di coerenza, che questo limite proiettivo esiste come misura di probabilità su uno spazio prodotto appropriato: la torre bayesiana infinita ha un oggetto limite ben definito. Questo oggetto limite è la struttura formale del "conoscitore ideale" — il sistema che ha una rappresentazione completa e coerente di tutti i livelli della propria incertezza su se stesso.

La metacognizione reale — quella di un sistema cognitivo finito — è una *troncatura finita* di questa torre: il sistema mantiene rappresentazioni esplicite di Θ attraverso π^k_t per $k \leq K$, dove K è la profondità metacognitiva del sistema, e collassa implicitamente tutti i livelli superiori in un prior fisso. La profondità metacognitiva K è una proprietà del sistema — non tutti i sistemi cognitivi hanno la stessa capacità di rappresentazione riflessiva — e la sua variazione produce differenze qualitative nel comportamento.

I.3 L'aggiornamento bayesiano in presenza di auto-osservazione

La dinamica della torre bayesiana gerarchica è governata dall'aggiornamento di Bayes a ogni livello. Al livello base, il sistema aggiorna il proprio prior π^0_t sui parametri cognitivi in risposta alle osservazioni:

$$\pi^0_{t+1}(\theta) \propto P(o_{t+1} \mid \theta) \cdot \pi^0_t(\theta)$$

dove $P(o_{t+1} \mid \theta)$ è la verosimiglianza dell'osservazione o_{t+1} dato che i parametri cognitivi siano θ . Al livello meta, il sistema aggiorna il prior di secondo ordine π^1_t in risposta alla divergenza tra la previsione del meta-modello e il prior risultante:

$$\pi^1_{t+1}(\pi^0) \propto P(\pi^0_{t+1} \mid \pi^0) \cdot \pi^1_t(\pi^0)$$

dove $P(\pi^0_{t+1} \mid \pi^0)$ è la verosimiglianza del prior aggiornato dato il meta-prior π^0 . Questa struttura si propaga ricorsivamente a tutti i livelli della torre.

Il problema fondamentale dell'auto-osservazione emerge qui con chiarezza. Le osservazioni $O_{0:t}$ disponibili al sistema per aggiornare il proprio prior includono le proprie uscite comportamentali, gli stati accessibili alla coscienza, e le conseguenze delle proprie azioni nel mondo. Ma il sistema non può osservare direttamente la propria struttura parametrica θ : può solo osservare le proiezioni di θ attraverso la funzione di uscita. Formalmente:

$$o_t = h(\theta, s_t, \epsilon_t)$$

dove h è la funzione di uscita che mappa lo stato interno s_t e il rumore ϵ_t nell'osservazione o_t attraverso i parametri θ . Il problema di inferire θ da o_t è un problema di identificazione che, in generale, è mal posto: più strutture parametriche diverse $\theta_1 \neq \theta_2$ possono produrre la stessa distribuzione di osservazioni, rendendo l'auto-conoscenza parametrica fondamentalmente non identificabile.

I.4 Non identificabilità e auto-opacità strutturale

La non identificabilità del problema di auto-inferenza ha struttura precisa. Si definisce la classe di equivalenza di due stati parametrici:

$$\theta_1 \sim \theta_2 \Leftrightarrow \forall s, P(O \mid \theta_1, s) = P(O \mid \theta_2, s)$$

Due stati parametrici sono osservazionalmente equivalenti se producono la stessa distribuzione di osservazioni per qualsiasi stato del mondo. La partizione di Θ in classi di equivalenza $\sim$ è la struttura dell'auto-opacità: il sistema non può distinguere tra elementi della stessa classe di equivalenza sulla sola base delle proprie osservazioni.

La dimensionalità dello spazio quoziente $\Theta/\sim$ — il numero di parametri identificabili — è il grado effettivo di auto-trasparenza del sistema. Per sistemi cognitivi con parametri altamente ridondanti e funzioni di uscita a bassa dimensionalità — che è esattamente il caso dei substrati neurali ad alta dimensione e degli output comportamentali a bassa dimensione — la dimensionalità di $\Theta/\sim$ può essere molto inferiore a quella di Θ : la maggior parte della struttura parametrica interna è nella penombra dell'auto-opacità.

Formalmente, la dimensionalità del sottospazio identificabile è bounded dalla capacità del canale di auto-osservazione:

$$\dim(\Theta/\sim) \leq I(\Theta ; O) / \log 2$$

dove $I(\Theta ; O)$ è l'informazione mutua tra i parametri cognitivi e le osservazioni. Questa quantità — la capacità del canale di auto-osservazione — è il tetto superiore dell'auto-conoscenza identificabile. Per il sistema cognitivo umano, studi sulla latenza delle risposte metacognitive e sulla calibrazione dei giudizi di confidenza suggeriscono che $I(\Theta ; O)$ è dell'ordine di 10^2 - 10^3 bit — una frazione minima dell'entropia totale $H(\Theta)$ del substrato, che nel quadro della seconda monografia è stimata in 10^4 - 10^5 bit per domini cognitivi complessi.

I.5 Il prior metacognitivo come iperparametro e il problema della sua specificazione

Nell'approccio bayesiano gerarchico, il prior metacognitivo $\pi^0(\theta)$ — la distribuzione di credenza iniziale del sistema sui propri parametri cognitivi — è esso stesso un iperparametro che deve essere specificato prima dell'aggiornamento. Ma come viene specificato il prior metacognitivo? Mediante un iper-prior $\pi^1(\pi^0)$: la distribuzione di secondo ordine. E come viene specificato l'iper-prior? Mediante un iper-iper-prior $\pi^2(\pi^1)$. E così la regressione sale la torre fino al limite proiettivo Π^∞ .

Questa regressione non è circolare — è la struttura propria di qualsiasi modello bayesiano gerarchico — ma produce un problema di specificazione che non ha soluzione pienamente soddisfacente. Il sistema deve, prima o poi, trancare la torre e

fissare un prior al livello $k = K$ senza averlo derivato da un livello superiore: deve accettare un prior non giustificato dalla propria struttura interna. Questo è il punto in cui l'architettura cognitiva incontra un vincolo che non è riducibile a incertezza di secondo ordine: il bordo superiore della torre bayesiana è un dato strutturale del sistema, non una scelta.

Per i sistemi cognitivi biologici, il prior metacognitivo di alto livello è in parte filogeneticamente determinato — le strutture neurali che realizzano le aspettative di base sulla propria competenza e capacità sono costruite nel corso dello sviluppo sotto vincoli genetici — e in parte costruito attraverso l'esperienza. La dinamica del sistema accoppiato substrato-linguaggio della seconda monografia include, come caso particolare, la costruzione del prior metacognitivo attraverso la co-evoluzione: il termine f_{lang} nell'equazione del substrato aggiorna non solo le strutture di primo ordine ma anche quelle metacognitive, orientando la geometria di $\mathbb{R}^n$ verso le strutture descritte dal vocabolario metacognitivo della propria lingua e cultura.

I.6 Calibrazione e deviazioni sistematiche dalla calibrazione ottimale

Un sistema bayesiano ideale produce stime probabilistiche calibrate: la probabilità soggettiva che il sistema assegna a un evento corrisponde alla frequenza relativa con cui quell'evento si verifica nella classe degli eventi a cui il sistema assegna quella probabilità. La calibrazione è la proprietà di allineamento tra il prior metacognitivo e la struttura effettiva del sistema.

Per i sistemi cognitivi reali, la calibrazione metacognitiva è sistematicamente imperfetta, e le deviazioni dalla calibrazione ottimale hanno struttura non casuale. Le deviazioni documentate includono:

L'overconfidence nei domini di competenza media: i soggetti sovrastimano sistematicamente la propria accuratezza nei compiti di media difficoltà. Formalmente, questo corrisponde a un prior metacognitivo $\pi^0(\theta)$ che sovrastima la norma del parametro θ nel dominio rilevante: il sistema crede di avere più struttura cognitiva nel dominio di quanto non ne abbia effettivamente.

L'underconfidence nei domini di competenza alta: i soggetti sottostimano sistematicamente la propria accuratezza nei compiti facili, un effetto noto come effetto Kruger inverso. Formalmente, questo corrisponde a un prior metacognitivo che non ha aggiornato adeguatamente verso l'alto la stima della propria competenza in un dominio dove l'expertise è alta ma la sua struttura è in $\mathbb{R}^n$ anziché in L — dove la competenza è procedurale e opaca all'auto-osservazione proposizionale.

Il *blindspot bias*: la tendenza sistematica a riconoscere i bias cognitivi negli altri ma non in se stessi. Questo è formalmente il caso in cui il prior metacognitivo del sistema è calibrato sulla classe di altri agenti (ha accesso alle osservazioni degli errori altrui) ma non sulla propria struttura interna (non ha accesso all'equivalente delle osservazioni sulle proprie performance dal punto di vista esterno).

La struttura di queste deviazioni è coerente con il modello della torre bayesiana troncata: un sistema che ha profondità metacognitiva K finita produce sistematicamente errori al livello K , dove la torre viene troncata e un prior non aggiornato viene fissato. Le deviazioni dalla calibrazione ottimale sono la traccia comportamentale della troncatura della torre.

I.7 La struttura dell'aggiornamento metacognitivo ottimale

Si definisce l'aggiornamento metacognitivo ottimale come l'aggiornamento bayesiano che minimizza la perdita attesa del prior metacognitivo rispetto alla struttura vera del sistema. La funzione di perdita appropriata per l'aggiornamento metacognitivo è la divergenza di Kullback-Leibler:

$$\mathcal{L}(\pi^0, \theta^*) = D_{\text{KL}}(\delta_{\{\theta^*\}} \parallel \pi^0) = -\log \pi^0(\theta^*)$$

dove θ^* è il valore vero dei parametri cognitivi. L'aggiornamento ottimale minimizza l'aspettativa di questa perdita sotto la distribuzione vera di θ^* :

$$\pi^{0*} = \operatorname{argmin}_{\{\pi^0\}} \mathbb{E}_{\{\theta^* \sim P_{\text{true}}\}} [D_{\text{KL}}(\delta_{\{\theta^*\}} \parallel \pi^0)] = P_{\text{true}}$$

Il prior metacognitivo ottimale è la distribuzione vera dei propri parametri cognitivi — ovviamente. Il punto non è questa tautologia ma il corollario: qualsiasi prior $\pi^0 \neq P_{\text{true}}$ produce una perdita metacognitiva positiva:

$$\Delta\mathcal{L} = D_{\text{KL}}(P_{\text{true}} \parallel \pi^0) > 0$$

e questa perdita si propaga alle decisioni di allocazione delle risorse cognitive, producendo inefficienze che possono essere stimate e in linea di principio misurate. La teoria dell'allocazione ottimale delle risorse cognitive sotto auto-opacità — sviluppata nel Capitolo IV — prende questa perdita come il suo input fondamentale.

Capitolo II — I teoremi di impossibilità per l'auto-conoscenza: da Gödel a Lawvere

II.1 La galassia dei teoremi di impossibilità

Nel Novecento, la logica matematica ha prodotto una costellazione di risultati negativi — teoremi che stabiliscono che certi sistemi non possono fare certe cose — che appaiono a prima vista eterogenei ma che condividono una struttura profonda. Il teorema di incompletezza di Gödel (1931), il teorema di indefinibilità della verità di Tarski (1936), il teorema di Löb (1955), il teorema di Rice (1953), il problema della fermata di Turing (1936), il teorema di incompletezza algoritmica di Chaitin

(1966-1975), il teorema di no-free-lunch di Wolpert e Macready (1997): ciascuno di questi risultati afferma che un sistema non può fare qualcosa rispetto a se stesso o alla propria classe. Ciascuno è stato interpretato in modo isolato, spesso strumentalizzato filosoficamente, raramente connesso agli altri mediante un framework unificante.

Il framework unificante esiste, ed è il teorema di punto fisso di Lawvere nella categoria della teoria degli insiemi, nella sua generalizzazione al contesto dei topos (Lawvere, 1969). Una volta che si riconosce che Gödel, Tarski, Rice e il problema della fermata sono tutti istanze del medesimo schema diagonale nella categoria appropriata, si ottiene non solo una visione unificata ma anche una previsione precisa: *qualsiasi* sistema che soddisfa certe condizioni di espressività e auto-riferimento è soggetto a un risultato di impossibilità della stessa famiglia. Per i sistemi cognitivi che modellano se stessi, questo non è una curiosità logica ma un vincolo architettonale.

II.2 Lo schema diagonale e il teorema di Lawvere

Il teorema di punto fisso di Lawvere è il risultato più generale della famiglia. Per formularlo in modo accessibile senza sacrificarne la precisione, si parte dalla versione insiemistica.

Siano A e B due insiemi e sia $\varphi : A \rightarrow B^A$ una funzione che a ogni elemento $a \in A$ associa una funzione $\varphi_a : A \rightarrow B$. Il teorema afferma:

Teorema di Lawvere: Se φ è suriettiva — se cioè per ogni funzione $g : A \rightarrow B$ esiste $a \in A$ tale che $\varphi_a = g$ — allora ogni funzione $f : B \rightarrow B$ ha un punto fisso: $\exists b \in B$ tale che $f(b) = b$.

La dimostrazione è l'argomento diagonale nella sua forma più pura. Si definisce la funzione diagonale:

$$d : A \rightarrow B, \quad d(a) = f(\varphi_a(a))$$

Poiché φ è suriettiva, esiste $a^* \in A$ tale che $\varphi_{a^*} = d$. Quindi:

$$\varphi_{a^*}(a^*) = d(a^*) = f(\varphi_{a^*}(a^*))$$

che è esattamente la condizione di punto fisso per f con valore $b^* = \varphi_{a^*}(a^*)$.

La potenza del teorema di Lawvere è che il suo contraposizionale produce immediatamente tutti i risultati di impossibilità della famiglia. Se si vuole che una certa funzione $f : B \rightarrow B$ non abbia punti fissi — per esempio, la funzione di negazione $f(v) = \neg v$ nel caso booleano — allora la premessa della suriettività di φ deve essere falsa: non può esistere una

funzione suriettiva da A a B^A , il che significa che A non può "nominare" o "rappresentare" tutte le funzioni da A a B .

II.3 Il teorema di Gödel come istanza di Lawvere

Il teorema di incompletezza di Gödel si ottiene prendendo $A = \mathbb{N}$ (i numeri naturali, che codificano le formule del sistema formale), $B = \{0,1\}$ (i valori di verità), $f = \neg$ (la negazione), e $\varphi : \mathbb{N} \rightarrow \{0,1\}^{\mathbb{N}}$ la funzione che a ogni numero n associa la funzione di prova parziale del sistema: $\varphi_n(m) = 1$ se n è il codice Gödel di un teorema che afferma la dimostrabilità del numero m , 0 altrimenti.

Se il sistema formale fosse completo — se cioè ogni formula vera fosse dimostrabile — allora φ sarebbe suriettiva: per ogni funzione $g : \mathbb{N} \rightarrow \{0,1\}$ ci sarebbe un numero n che la "dimostra". Il teorema di Lawvere dice allora che la negazione $\neg$ avrebbe un punto fisso: una formula che equivale alla propria negazione. Ma questo è una contraddizione (salvo che il sistema sia inconsistente). Quindi φ non è suriettiva: esiste una funzione $g : \mathbb{N} \rightarrow \{0,1\}$ che nessun numero del sistema "dimostra" — c'è una formula vera che il sistema non può dimostrare.

La formula di Gödel G — "questa formula non è dimostrabile" — è esattamente il punto fisso diagonale $d(a^*) = f(\varphi_{a^*}(a^*))$ costruito nella dimostrazione del teorema di Lawvere. Non è un'anomalia tecnica: è la struttura della diagonalizzazione che compare ogni volta che un sistema tenta di rappresentare completamente se stesso.

II.4 Il teorema di Tarski come istanza di Lawvere

Il teorema di indefinibilità della verità di Tarski afferma che in nessun sistema formale coerente e sufficientemente espressivo esiste un predicato $V(x)$ tale che $V(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ per ogni formula φ — nessun predicato che catturi la verità delle formule del sistema all'interno del sistema stesso.

Nell'istanza di Lawvere: $A = \mathbb{N}$, $B = \{0,1\}$, $f = \neg$, $\varphi_n(m) = V(\ulcorner P(n,m) \urcorner)$ dove $P(n,m)$ è la formula " n soddisfa m ". Se V fosse definibile nel sistema e catturasse la verità, φ sarebbe suriettiva, e il teorema di Lawvere produrrebbe un punto fisso per $\neg$ — una formula equivalente alla propria falsità — contraddizione. Quindi V non è definibile.

Il contenuto cognitivo del teorema di Tarski è preciso: nessun sistema può contenere internamente una rappresentazione completa e fedele della propria struttura veritativa. Il sistema può parlare della verità di asserzioni su altri sistemi — Peano Arithmetic può definire la verità dell'aritmetica del grado inferiore — ma non della propria. Ogni tentativo di auto-rappresentazione veritativa completa produce o inconsistenza o incompletezza.

II.5 Il teorema di Löb e il limite dell'auto-fiducia

Il teorema di Löb (1955) è meno noto dei suoi fratelli ma ha importanza diretta per la metacognizione. Il teorema afferma:

Teorema di Löb: In qualsiasi sistema formale P che contenga l'aritmetica di Peano, per qualsiasi formula φ :

$P \vdash \Box_P \varphi \rightarrow \varphi$ implica $P \vdash \varphi$

dove $\Box_P \varphi$ è la formula " φ è dimostrabile in P ". In altre parole: se il sistema può dimostrare che "se φ è dimostrabile allora φ è vera", allora il sistema può già dimostrare φ direttamente.

Il teorema di Löb è il limite dell'auto-fiducia del sistema. Un sistema ragionevole dovrebbe essere in grado di dimostrarsi che "se posso dimostrare φ , allora φ è vera" — questa sembra una proprietà minimale della propria affidabilità. Ma il teorema di Löb dice che questa capacità è disponibile soltanto per le formule che il sistema può già dimostrare: il sistema non può acquisire nuova conoscenza attraverso la propria fiducia in se stesso.

Il contenuto metacognitivo è: l'auto-validazione non genera conoscenza. Un sistema che giustifica le proprie credenze attraverso la fiducia nella propria affidabilità ("credo p perché mi fido del mio giudizio, e il mio giudizio dice p ") non acquisisce una giustificazione epistemica aggiuntiva per p al di là di quella già disponibile. L'auto-fiducia è epistemicamente sterile: il teorema di Löb formalizza l'intuizione del motto "haec stultitia facit" — credere di sapere perché si crede di sapere è la struttura della pseudoconoscenza, non della conoscenza.

Nel framework bayesiano gerarchico, il teorema di Löb corrisponde alla non contrattività della mappa di aggiornamento: il prior metacognitivo non può auto-validarsi producendo conoscenza aggiuntiva. L'aggiornamento bayesiano produce conoscenza soltanto in risposta a osservazioni esterne — non all'auto-riflessione sul proprio prior.

II.6 Il teorema di Rice e l'opacità delle proprietà semantiche

Il teorema di Rice (1953) afferma che nessuna proprietà semantica non banale dei programmi è decidibile. Formalmente: per qualsiasi proprietà P dei programmi che sia non banale — che sia vera per alcuni programmi e falsa per altri — e semantica — che dipenda dal comportamento del programma anziché dalla sua sintassi — non esiste un algoritmo che, dato un programma arbitrario q , decida se q ha la proprietà P .

Questo teorema è un'estensione del problema della fermata (che è il caso particolare in cui P è la proprietà di terminare) e si ottiene da esso per riduzione. Nel quadro di Lawvere: A è l'insieme dei programmi (codificati come numeri naturali), $B = \{0,1\}$, $f = \neg$, e $\varphi_q(r) = P(q[r])$ dove $q[r]$ è il programma q applicato all'input r . La suriettività di φ significherebbe che ogni funzione booleana su programmi è realizzata da un programma che la computa — il che è vero per la semantica operativa completa — e il teorema di Lawvere produce un punto fisso per $\neg$, che contraddice la decidibilità.

Il contenuto cognitivo del teorema di Rice per la metacognizione è radicale: nessun sistema cognitivo può determinare algebricamente, su se stesso, se possiede una proprietà semantica non banale. Nessuna procedura di auto-ispezione può determinare se "capisco davvero questo concetto" anziché "produco l'output corretto per le istanze di training viste finora". La distinzione tra comprensione genuina e simulazione comportamentale fedele è, per il teorema di Rice, indecidibile dall'interno. Questo vale per i modelli di linguaggio — che non possono distinguere dall'interno tra grounding semantico e pattern matching statistico — ma vale altrettanto per i sistemi cognitivi umani, che non possono determinare attraverso l'auto-ispezione se la propria intuizione esperta è genuina competenza procedurale o ritualizzazione comportamentale.

II.7 Il teorema di Chaitin e il limite dell'auto-descrizione compressa

Il teorema di incompletezza algoritmica di Chaitin fornisce una versione quantitativa dei risultati di impossibilità. Sia $KC(x)$ la complessità di Kolmogorov di una stringa x — la lunghezza del programma più corto che la genera. Il teorema di Chaitin afferma:

Teorema di Chaitin: Per qualsiasi sistema formale F con assiomi di complessità $KC(F) = k$, F non può dimostrare che $KC(x) > k + c$ per nessuna stringa x , dove c è una costante che dipende solo da F .

In altre parole: un sistema non può dimostrare che nessun oggetto è più complesso del sistema stesso più una costante. Il sistema non può riconoscere complessità superiore alla propria all'interno della propria logica.

Il corollario per l'auto-descrizione è immediato: un sistema non può produrre una descrizione di se stesso che sia più corta del sistema stesso (per la definizione di complessità di Kolmogorov, la descrizione più corta di x ha lunghezza $KC(x)$) ma non può nemmeno dimostrare che la propria auto-descrizione è ottimalmente compressa. La meta-compressione — la compressione della propria struttura compressa — è soggetta allo stesso limite: il sistema non può certificare l'ottimalità della propria auto-rappresentazione.

Nella torre bayesiana gerarchica, il teorema di Chaitin corrisponde al seguente bound:

$$KC(\pi^0) \geq KC(\theta^*) - c$$

la complessità del prior metacognitivo non può essere inferiore a quella dei parametri veri meno una costante. Un prior metacognitivo "troppo semplice" — una distribuzione con bassa complessità algoritmica — non può catturare fedelmente parametri cognitivi ad alta complessità. La semplicità del meta-modello è un limite alla sua fedeltà, e questo limite è precisamente quantificato dalla differenza $KC(\theta^*) - KC(\pi^0)$.

II.8 Il teorema di no-free-lunch come limite all'ottimizzazione auto-diretta

Wolpert e Macready (1997) hanno dimostrato il teorema di no-free-lunch per l'ottimizzazione: per qualsiasi coppia di algoritmi di ottimizzazione A_1 e A_2 , mediando su tutti i possibili problemi di ottimizzazione f con la stessa misura, le prestazioni attese di A_1 e A_2 sono identiche. In altre parole: nessun algoritmo di ottimizzazione è universalmente superiore agli altri — ogni vantaggio su una classe di problemi è compensato da uno svantaggio su un'altra classe.

Nel contesto della metacognizione, il teorema di no-free-lunch si applica all'auto-ottimizzazione cognitiva: la scelta di quale strategia metacognitiva adottare — quale meta-modello costruire, quale procedura di allocazione delle risorse usare — non può essere ottimizzata universalmente. Un meta-modello che funziona bene per un certo tipo di problemi cognitivi inevitabilmente funziona meno bene per altri. La scelta del meta-modello è essa stessa una scommessa: un prior metacognitivo implicito sulla distribuzione dei problemi che il sistema incontrerà.

Formalmente, sia M lo spazio dei meta-modelli metacognitivi e sia $U(m, f)$ l'utilità del meta-modello m sul problema cognitivo f . Il teorema di no-free-lunch stabilisce:

$$\sum_f U(m_1, f) = \sum_f U(m_2, f) \text{ per ogni } m_1, m_2 \in M$$

dove la somma è su tutti i possibili problemi cognitivi con la stessa misura. Il meta-modello ottimale non esiste in assoluto — esiste solo relativo alla distribuzione $P(f)$ dei problemi che il sistema si aspetta di incontrare. L'adattamento del meta-modello all'ambiente cognitivo specifico dell'agente è la struttura formale di ciò che colloquialmente chiamiamo "saggezza pratica": la capacità di scegliere la strategia metacognitiva appropriata al contesto anziché la strategia ottimale in astratto.

II.9 Il teorema di Lawvere come schema unificante: la mappa dell'impossibilità

Si riassume ora la tassonomia dei teoremi di impossibilità per l'auto-conoscenza attraverso lo schema di Lawvere, che li unifica nella struttura dell'argomento diagonale nella categoria appropriata.

Ogni risultato ha la forma: *se un sistema può rappresentare completamente (in senso da precisare) se stesso, allora esiste un oggetto o una proprietà che quel sistema non può gestire (dimostrare, decidere, comprimere, ottimizzare)*. La struttura è sempre la stessa: l'auto-rappresentazione completa genera una diagonalizzazione che produce un oggetto fuori dall'alcance del sistema.

La tassonomia è:

Gödel (1931): sistema formale, auto-rappresentazione = enumerazione delle dimostrazioni, impossibilità = completezza. L'auto-enumerazione completa delle dimostrazioni genera formule non dimostrabili.

Tarski (1936): sistema formale, auto-rappresentazione = predicato di verità, impossibilità = definibilità interna della verità. L'auto-predicazione veritativa completa genera paradosso del mentitore formalizzato.

Turing (1936): macchine di Turing, auto-rappresentazione = enumerazione dei programmi, impossibilità = decidibilità della terminazione. L'auto-enumerazione completa dei programmi genera il problema della fermata.

Löb (1955): sistema formale, auto-rappresentazione = predicato di dimostrabilità, impossibilità = auto-validazione epistemicamente generativa. L'auto-fiducia non produce conoscenza nuova.

Rice (1953): programmi, auto-rappresentazione = proprietà semantiche, impossibilità = decidibilità di qualsiasi proprietà semantica non banale. L'auto-ispezione semantica completa è impossibile.

Chaitin (1966-75): stringhe, auto-rappresentazione = auto-descrizione compressa, impossibilità = certificazione dell'ottimalità della propria auto-descrizione. Il sistema non può riconoscere complessità superiore alla propria.

No-free-lunch (1997): algoritmi di ottimizzazione, auto-rappresentazione = meta-strategia universale, impossibilità = ottimizzazione universale. Nessun meta-modello è universalmente ottimale.

La lezione unificata di questa tassonomia per la metacognizione è che l'auto-conoscenza completa è impossibile in ogni senso formalmente preciso in cui la si voglia definire. Questo non è pessimismo epistémico ma vincolo architettuale: il sistema che tenta di auto-modellarsi completamente genera invariabilmente strutture che sfuggono al proprio meta-modello. La metacognizione competente non è quella che aspira all'auto-conoscenza completa — aspirazione impossibile — ma quella che ha un modello calibrato della propria incompletezza: che sa, per i diversi domini cognitivi, quanto del proprio substrato è fuori dall'alcance della propria auto-rappresentazione, e agisce di conseguenza.

II.10 Implicazioni dirette per l'architettura del meta-modello

I teoremi di impossibilità hanno implicazioni dirette per l'architettura del meta-modello — per come il sistema metacognitivo deve essere costruito per essere efficace sotto i vincoli che i teoremi impongono.

La prima implicazione è che il meta-modello non deve puntare alla completezza: l'auto-rappresentazione completa è formalmente impossibile, e un meta-modello che ci tenta produce paradosso o circolarità. Il meta-modello efficace è quello che identifica i sottospazi di Θ identificabili attraverso l'auto-osservazione — le

dimensioni di $\Theta/\sim$ — e si concentra su quelli, lasciando il resto coperto da priors vaghi che riflettono l'auto-opacità genuina.

La seconda implicazione è che il meta-modello deve modellare esplicitamente la propria incompletezza: deve avere una stima della dimensione dello spazio $\Theta/\sim$ rispetto all'intero Θ , e usare questa stima per calibrare la propria confidenza nelle stime metacognitive. Un meta-modello che non sa di non sapere — che ha piena confidenza nelle proprie stime su dimensioni di Θ che sono fuori dalla partizione $\Theta/\sim$ — produce errori sistematici di auto-calibrazione.

La terza implicazione è che il meta-modello deve essere aggiornabile: i teoremi di impossibilità valgono per qualsiasi specifico meta-modello, ma non impediscono al sistema di avere meta-modelli successivamente migliori. L'auto-conoscenza non è completabile, ma è migliorabile iterativamente — ogni ciclo di osservazione, aggiornamento bayesiano e revisione del meta-modello riduce la distorsione metacognitiva nel sottoinsieme dei parametri raggiungibili attraverso l'auto-osservazione.

La quarta implicazione riguarda il sistema multi-agente. Nei sistemi di agenti in interazione — organizzazioni, comunità epistemiche, collettivi sociali — la metacognizione collettiva incontra teoremi di impossibilità aggiuntivi che si sovrappongono a quelli individuali. Il teorema di impossibilità di Arrow per l'aggregazione delle preferenze è, come si vedrà nel Capitolo V, un'istanza della medesima struttura diagonale applicata non ai sistemi singoli ma alle funzioni di aggregazione collettiva: anch'esso afferma che nessuna procedura di aggregazione può soddisfare simultaneamente un insieme di proprietà minimali ragionevoli, e la sua connessione con i teoremi di auto-conoscenza individuale è la struttura formale del Capitolo V.

Capitolo III — La metacognizione come compressione di secondo ordine

III.1 Il problema della compressione gerarchica

La seconda monografia del ciclo ha stabilito che la codifica linguistica del substrato è una compressione lossy con una funzione di tasso-distorsione $R_{\text{sub}}(D)$ che misura il costo informativo di ogni livello di distorsione. La metacognizione introduce una struttura analoga ma di ordine superiore: il meta-modello $\pi^0(\theta)$ è una compressione del sistema cognitivo θ^* , e questa compressione è anch'essa lossy, con la propria funzione di tasso-distorsione $R_{\text{meta}}(D)$.

La struttura a due livelli è:

$$\begin{aligned}\theta^* &\xrightarrow{[\Phi_{\text{sub}}]} m_{\text{lang}} && \text{(compressione primo ordine: substrato} \rightarrow \text{linguaggio)} \\ \theta^* &\xrightarrow{[\Phi_{\text{meta}}]} \pi^0 && \text{(compressione secondo ordine: parametri} \rightarrow \text{meta-modello)}\end{aligned}$$

Le due compressioni sono formalmente analoghe ma operano su spazi diversi e con funzioni di distorsione diverse. La compressione di primo ordine lavora in $\mathbb{R}^n$ e usa la distanza angolare. La compressione di secondo ordine lavora nello spazio dei parametri Θ e usa la divergenza di Kullback-Leibler come misura di distorsione:

$$d_{\text{meta}}(\theta^*, \pi^0) = D_{\text{KL}}(\delta_{\{\theta^*\}} \parallel \pi^0) = -\log \pi^0(\theta^*)$$

la log-perdita del meta-modello sul valore vero dei parametri.

III.2 La funzione di tasso-distorsione metacognitiva

Si definisce la funzione di tasso-distorsione metacognitiva come il minimo dell'informazione trasmessa dal meta-modello compatibile con una distorsione media al più D :

$$R_{\text{meta}}(D) = \min_{\{q(\pi^0 \mid \theta^*) : \mathbb{E}[d_{\text{meta}}] \leq D\}} I(\Theta^*; \Pi^0)$$

dove $q(\pi^0 \mid \theta^*)$ è la distribuzione condizionale del meta-modello dato il valore vero dei parametri, e $I(\Theta^*; \Pi^0)$ è l'informazione mutua tra parametri veri e meta-modello.

Per una sorgente gaussiana su Θ^* con covarianza Σ_{θ} , la funzione di tasso-distorsione metacognitiva ha la forma:

$$R_{\text{meta}}(D) = (1/2) \cdot \sum_i \max(0, \log(\lambda_i/D))$$

dove λ_i sono gli autovalori della matrice di covarianza Σ_{θ} . Questa è la soluzione water-filling di Shannon: la capacità di trasmissione viene allocata preferibilmente alle direzioni dello spazio dei parametri con alta varianza (alta λ_i), e le direzioni con varianza inferiore alla soglia D vengono eliminate dal meta-modello.

Il risultato water-filling ha un'interpretazione cognitiva precisa: il meta-modello ottimale alloca la propria capacità di rappresentazione preferenzialmente alle dimensioni dello spazio dei parametri cognitivi con maggiore varianza — le dimensioni in cui il sistema ha maggiore incertezza su se stesso. Le dimensioni ben note del proprio funzionamento non necessitano di modellazione esplicita nel meta-modello; le dimensioni più opache e variabili richiedono allocazione prioritaria.

III.3 L'interazione tra le due compressioni: distorsione composta

La compressione di primo ordine (substrato $\rightarrow$ linguaggio) e la compressione di secondo ordine (parametri $\rightarrow$ meta-modello) non sono indipendenti. La qualità del meta-modello sulla competenza linguistica — quanto accuratamente π^0 stima i parametri m_lang del sistema — influenza direttamente la qualità della codifica di primo ordine: un meta-modello che sovrastima la propria competenza linguistica in un dominio produrrà codifiche di primo ordine sub-ottimali in quel dominio, perché il sistema non alloca risorse di riflessione e revisione dove sarebbero necessarie.

Si definisce la distorsione composita come la composizione delle due perdite:

$$d_comp(\theta^*, \pi^0) = d_meta(\theta^*, \pi^0) + \mathbb{E}_{\{\theta \sim \pi^0\}}[d_sub(v, \Phi(v))]$$

la somma della perdita metacognitiva e della perdita attesa di codifica di primo ordine sotto il meta-modello. La funzione di tasso-distorsione composita:

$$R_comp(D) = \min_{\{q, \Phi : \mathbb{E}[d_comp] \leq D\}} I(\Theta^* ; \Pi^0) + I(V ; \Phi(V))$$

è la quantità minima totale di informazione che deve essere trasmessa, rispettivamente dal meta-modello e dalla codifica di primo ordine, per ottenere una distorsione composita al più D . Questa funzione ha un'implicazione fondamentale: esiste un punto ottimale di bilanciamento tra la capacità allocata al meta-modello e quella allocata alla codifica di primo ordine. Investire tutta la capacità computazionale nella raffinatezza del meta-modello lascia troppo poco per la codifica di primo ordine; viceversa, trascurare il meta-modello produce allocazioni sub-ottimali della capacità di primo ordine.

III.4 Il meta-modello come struttura di controllo della compressione

Nel framework della compressione gerarchica, il meta-modello non è soltanto una rappresentazione passiva dello stato del sistema: è la struttura che controlla il processo di compressione di primo ordine. Il meta-modello determina dove la funzione di codifica Φ viene applicata con alta risoluzione e dove viene applicata con risoluzione bassa — in quali domini il sistema investe risorse computazionali di elaborazione esplicita e in quali si affida alla risposta automatica del substrato.

Formalmente, si modella questo come un problema di selezione della mappa di codifica. Il sistema ha a disposizione una famiglia di funzioni di codifica $\{\Phi_\theta : \theta \in \Theta\}$, ciascuna adattata a un certo sottoinsieme dello spazio degli stati cognitivi. Il meta-modello $\pi^0(\theta)$ determina la distribuzione di probabilità sulla famiglia di codifiche:

$$P(\Phi = \Phi_\theta \mid \text{contesto } c) \propto \pi^0(\theta) \cdot P(c \mid \theta)$$

la codifica attivata per un dato contesto è selezionata bayesianamente dal meta-modello. Un meta-modello accurato seleziona la codifica appropriata al contesto; un meta-modello distorto seleziona sistematicamente codifiche sub-ottimali.

Il caso patologico è il meta-modello "troppo piatto" — un prior metacognitivo quasi uniforme su Θ che non discrimina tra domini diversi. Un sistema con questo tipo di meta-modello usa la stessa strategia di codifica per tutti i contesti, ignorando le differenze strutturali tra domini. Questo corrisponde all'incapacità di passare tra pensiero analitico esplicito e pensiero intuitivo automatico in modo appropriato al tipo di problema — una forma di rigidità metacognitiva che il framework quantifica precisamente come la divergenza tra il prior metacognitivo π^0 e la distribuzione ottimale di selezione della codifica.

III.5 L'incertezza di secondo ordine come segnale di meta-comprensione

Nella torre bayesiana gerarchica, l'incertezza di secondo ordine — la varianza del meta-prior π^1 sulla distribuzione π^0 — è un segnale della qualità della meta-comprensione. Un sistema con alta incertezza di secondo ordine in un certo dominio sa che non sa — ha un meta-modello che registra la propria incalibrazione in quel dominio. Un sistema con bassa incertezza di secondo ordine ma meta-modello distorto non sa che non sa — è nel caso del Dunning-Kruger: alta confidenza metacognitiva accoppiata a bassa accuratezza di primo ordine.

Si definisce l'*epistemic gap* come la differenza tra l'incertezza di secondo ordine del sistema e l'errore di primo ordine atteso:

$$EG = \text{Var}_{\pi^1}(\pi^0) - \mathbb{E}_{\pi^1}[d_{\text{meta}}(\theta^*, \pi^0)]$$

Un epistemic gap positivo indica che il sistema è sovra-incerto rispetto al proprio errore atteso — conosce se stesso meno bene di quanto l'auto-osservazione consenta, suggerendo che non sta sfruttando adeguatamente le informazioni disponibili per l'aggiornamento metacognitivo. Un epistemic gap negativo indica che il sistema è sotto-incerto — ha più confidenza di quanta l'evidenza giustifichi. Il caso $EG = 0$ è la calibrazione ottimale di secondo ordine: il sistema ha esattamente l'incertezza corrispondente al proprio errore atteso.

Capitolo IV — Allocazione delle risorse cognitive sotto auto-opacità

IV.1 Il problema di allocazione come decisione sotto auto-ignoranza

I Capitoli precedenti hanno costruito il framework formale della metacognizione come compressione di secondo ordine e ne hanno identificato i limiti attraverso i teoremi di impossibilità. Questo capitolo affronta il problema pratico che la metacognizione è chiamata a risolvere: l'allocazione delle risorse cognitive limitate del sistema tra domini e compiti diversi, in condizioni di auto-opacità strutturale.

Le risorse cognitive sono limitate per ragioni sia biologiche — l'energia metabolica disponibile alla corteccia prefrontale è finite — sia computazionali — l'elaborazione esplicita è seriale e lenta rispetto all'elaborazione subpropositiva parallela. Il sistema deve allocare queste risorse in modo da massimizzare l'utilità attesa delle proprie uscite cognitive, tenendo conto che:

- (a) la propria competenza nei diversi domini è nota solo approssimativamente attraverso il meta-modello π^0 ;
- (b) il costo dell'elaborazione esplicita varia per dominio;
- (c) l'investimento in elaborazione esplicita modifica la propria competenza futura (attraverso il meccanismo di apprendimento del sistema dinamico accoppiato della seconda monografia).

IV.2 Il problema di ottimizzazione formale

Si formalizza il problema di allocazione. Il sistema ha un budget computazionale totale B da distribuire tra k domini $D_1, \dots, D_k$. Sia $b_i \geq 0$ la risorsa allocata al dominio D_i , con $\sum_i b_i \leq B$. L'utilità dell'elaborazione nel dominio D_i con risorsa b_i è:

$$U_i(b_i, \theta_i) = u_i \cdot (1 - \exp(-b_i / \tau(\theta_i)))$$

dove u_i è l'utilità massima del dominio D_i e $\tau(\theta_i)$ è la costante di tempo dell'apprendimento nel dominio — la quantità di risorsa necessaria per raggiungere metà del potenziale del dominio — che dipende dai parametri cognitivi del sistema per quel dominio.

Il sistema non conosce θ_i direttamente: conosce soltanto la propria distribuzione metacognitiva $\pi^0(\theta_i)$. Il problema di allocazione ottimale sotto auto-opacità è:

$$b^* = \operatorname{argmax}_{\{b : \sum_i b_i \leq B\}} \sum_i \mathbb{E}_{\{\theta_i \sim \pi^0\}} [U_i(b_i, \theta_i)]$$

Il problema è un problema di ottimizzazione stocastica la cui soluzione dipende dalla struttura del meta-modello π^0 . La soluzione ottimale esatta richiede di conoscere i momenti della distribuzione $\pi^0(\theta_i)$ per ogni dominio i .

IV.3 L'allocazione ottimale: la soluzione bayesiana

Per il caso in cui la funzione di utilità è concava in b_i — il che è garantito dalla struttura esponenziale della U_i scelta — la soluzione del problema di allocazione ottimale è determinata dalla condizione di Karush-Kuhn-Tucker:

$$\partial/\partial b_i \mathbb{E}_{\{\theta_i \sim \pi^0\}}[U_i(b_i, \theta_i)] = \lambda \text{ per ogni } i \text{ con } b_i > 0$$

dove λ è il moltiplicatore di Lagrange del vincolo di budget. Calcolando la derivata:

$$(u_i / \mathbb{E}[\tau_i]) \cdot \exp(-b_i / \mathbb{E}[\tau_i]) \cdot (1 + \text{Var}[\tau_i]/\mathbb{E}[\tau_i]^2 \cdot b_i/\mathbb{E}[\tau_i] + \dots) = \lambda$$

Per il caso in cui la varianza di τ_i sotto il meta-modello è piccola rispetto alla media, la soluzione approssimata è:

$$b_i^* \approx \mathbb{E}[\tau_i] \cdot \log(u_i / (\lambda \cdot \mathbb{E}[\tau_i]))$$

Questa soluzione ha un'interpretazione precisa: la risorsa ottimale allocata al dominio D_i è proporzionale alla costante di tempo attesa $\mathbb{E}[\tau_i]$ e al logaritmo del rapporto tra l'utilità del dominio e il costo marginale λ . I domini con alta utilità ricevono più risorse; i domini con bassa costante di tempo (apprendimento rapido) ricevono anch'essi più risorse perché offrono più rendimento per unità di risorsa investita.

Crucialmente, l'allocazione ottimale dipende dalla stima metacognitiva $\mathbb{E}[\tau_i]$ — la stima del sistema sulla propria velocità di apprendimento nel dominio i . Un meta-modello che sovrastima $\mathbb{E}[\tau_i]$ (stima di apprendere più lentamente di quanto si fa effettivamente) tende a sotto-allocare risorse al dominio, perdendo rendimento atteso. Un meta-modello che sottostima $\mathbb{E}[\tau_i]$ tende a sovra-allocare, sprecando risorse su un dominio già saturato. La perdita di utilità dovuta all'errore metacognitivo è:

$$\Delta U = U^*(b^*) - U(\tilde{b}^*)$$

dove b^* è l'allocazione ottimale sotto il meta-modello vero e $\tilde{b}^*$ è l'allocazione ottimale sotto il meta-modello distorto $\tilde{\pi}^0$. Questa perdita è il costo computazionale della miscalibrazione metacognitiva.

IV.4 La dimensione temporale: l'allocazione come investimento

Il problema di allocazione delle risorse cognitive ha una dimensione temporale che l'analisi statica del paragrafo precedente non cattura: investire risorse nell'apprendimento di un dominio oggi cambia la propria competenza futura, modificando sia i parametri cognitivi $\theta_i(t)$ sia il meta-modello $\pi^0(t)$. Il problema di allocazione ottimale su un orizzonte temporale $[0, T]$ è un problema di controllo ottimale stocastico.

Il sistema cognitivo è il controllore; il processo di apprendimento è il processo controllato; la variabile di controllo è l'allocazione $b_i(t)$ delle risorse al tempo t ; lo stato del sistema è $(\theta(t), \pi^0(t))$ — la competenza effettiva e il suo meta-modello. Il funzionale di utilità da massimizzare è:

$$J = \int_0^T e^{-\rho t} \cdot \sum_i U_i(b_i(t), \theta_i(t)) dt$$

dove ρ è il tasso di sconto temporale. Il principio del massimo di Pontryagin fornisce le condizioni di ottimalità:

$$\partial H / \partial b = 0, \quad -\partial H / \partial \theta = \mu, \quad -\partial H / \partial \pi^0 = v$$

dove H è l'Hamiltoniana del sistema e μ, v sono le variabili co-state associate rispettivamente alla competenza e al meta-modello. La variabile co-state v — il valore ombra del meta-modello — quantifica il valore marginale di un miglioramento nel meta-modello: quanto aumenterebbe l'utilità totale se il meta-modello fosse più accurato. Nei domini in cui v è alto, il sistema dovrebbe investire risorse non solo per apprendere direttamente ma anche per migliorare la propria auto-comprensione — per ridurre l'errore metacognitivo prima di procedere nell'apprendimento di primo ordine.

IV.5 La struttura dei punti fissi dell'allocazione: expertise come equilibrio

Un punto fisso dell'allocazione nel tempo è una configurazione (θ^*, b^*) in cui il sistema alloca le proprie risorse in modo da mantenere stabile la propria competenza: la competenza acquisita bilancia esattamente la competenza persa per decadimento sinaptico e obsolescenza. Questo punto fisso è l'equilibrio dell'expertise: la distribuzione stabile delle competenze che il sistema mantiene sotto un regime di apprendimento e decadimento costanti.

Per un sistema con k domini e velocità di decadimento δ_i per dominio, le condizioni di equilibrio sono:

$$b_i^* \cdot f_{\text{acq}}(\theta_i^*) = \delta_i \cdot \theta_i^* \quad \text{per ogni } i$$

dove f_{acq} è la funzione di acquisizione — la velocità di apprendimento in funzione delle risorse allocate. L'equilibrio richiede che la velocità di acquisizione bilanci la velocità di decadimento in ogni dominio.

Il punto fisso dell'allocazione ha la struttura di un attrattore del sistema di apprendimento: dato un budget B e una distribuzione di utilità $\{u_i\}$, il sistema converge a una distribuzione di competenza stabile indipendentemente dalle condizioni iniziali, purché il processo di aggiornamento del meta-modello sia sufficientemente accurato. La velocità di convergenza al punto fisso dipende dalla

qualità del meta-modello: un meta-modello accurato guida l'allocazione verso l'equilibrio più rapidamente, mentre un meta-modello distorto produce oscillazioni attorno all'equilibrio o convergenza a punti fissi sub-ottimali.

Capitolo V — Metacognizione collettiva e i teoremi di impossibilità per l'aggregazione

V.1 Dal sistema singolo al sistema multi-agente

I quattro capitoli precedenti hanno trattato la metacognizione come proprietà di un sistema cognitivo singolo. Ma la cognizione si svolge prevalentemente in contesti sociali — le decisioni vengono prese in collettivi, la conoscenza viene prodotta e validata in comunità epistemiche, le strutture di credenza si formano e si propagano in reti di agenti interagenti. La metacognizione collettiva — la capacità di un collettivo di agenti di avere un meta-modello della propria struttura cognitiva aggregata — è soggetta a teoremi di impossibilità che si sovrappongono a quelli individuali in modo strutturalmente rivelatore.

Il punto di connessione con i teoremi di impossibilità del Capitolo II è preciso: il teorema di Arrow per l'aggregazione delle preferenze e il teorema di Sen per l'impossibilità del liberale paretiano sono istanze della medesima struttura diagonale di Lawvere applicata non ai sistemi di auto-rappresentazione individuali ma alle funzioni di aggregazione collettiva. Riconoscere questa connessione unifica l'impossibilità dell'auto-conoscenza individuale con l'impossibilità dell'aggregazione collettiva in uno stesso schema formale: i limiti della cognizione sono limiti sia verso il basso (l'auto-opacità del substrato) sia verso l'alto (l'incoerenza dell'aggregazione).

V.2 Il teorema di Arrow come risultato di tipo Lawvere

Si considera un insieme $N = \{1, \dots, n\}$ di agenti, ciascuno con una relazione di preferenza $\succ_i$ sull'insieme X delle alternative. Una funzione di benessere sociale è una mappa $F : L(X)^n \rightarrow L(X)$ che a ogni profilo di preferenze individuali $(\succ_1, \dots, \succ_n)$ associa una preferenza collettiva $F(\succ_1, \dots, \succ_n)$.

Arrow (1951) dimostra che nessuna funzione di benessere sociale F soddisfa simultaneamente:

- (U) Dominio illimitato: F è definita su tutti i profili possibili.
- (P) Principio di Pareto: se tutti preferiscono x a y , la preferenza collettiva fa altrettanto.
- (IIA) Indipendenza dalle alternative irrilevanti: la preferenza collettiva tra x e y dipende solo dalle preferenze individuali tra x e y .
- (ND) Non-dittatura: nessun agente i la cui preferenza individuale determina sempre la collettiva.

La connessione con lo schema di Lawvere è la seguente. Si prenda $A = L(X)^n$ (lo spazio dei profili di preferenze), $B = L(X)$ (lo spazio delle preferenze collettive), e $\varphi : A \rightarrow B^A$ la funzione che a ogni profilo $(\succ_1, \dots, \succ_n)$ associa la funzione collettiva $F(\cdot)$ ristretta al contesto in cui uno degli agenti ha quella preferenza. La condizione (IIA) corrisponde alla struttura di fattorizzazione che rende φ candidata alla suriettività rispetto alle funzioni "locali". La condizione (ND) corrisponde alla non esistenza di un punto fisso "dittatoriale" per una certa famiglia di funzioni su B . Il teorema di Arrow è il risultato che la suriettività di φ (condizioni U+IIA) combinata con il principio P implica l'esistenza di un punto fisso — un dittatore — in contrasto con (ND).

La connessione formale è che sia il teorema di incompletezza di Gödel sia il teorema di impossibilità di Arrow sono istanze del teorema di Lawvere: in entrambi i casi, una struttura di auto-rappresentazione sufficientemente completa (nel caso di Gödel: un sistema che dimostra tutte le verità; nel caso di Arrow: una funzione di aggregazione che rispetta tutte le preferenze) genera un punto fisso impossibile (una formula vera ma non dimostrabile; un dittatore).

V.3 Il meta-modello collettivo e il problema della sua coerenza

Nel sistema multi-agente, ogni agente ha il proprio meta-modello $\pi^0_i(\theta_i)$ sulla propria architettura cognitiva individuale. Il meta-modello collettivo è una rappresentazione della struttura cognitiva del collettivo nel suo insieme — una distribuzione Π^0_C sullo spazio dei profili di parametri cognitivi collettivi $(\theta_1, \dots, \theta_n)$. Il problema della metacognizione collettiva è: come si forma Π^0_C a partire dai meta-modelli individuali π^0_i ?

La funzione di aggregazione metacognitiva $A : \{\pi^0_i\}_i \rightarrow \Pi^0_C$ è soggetta al teorema di Arrow nella sua versione per distribuzioni di probabilità. Hylland e Zeckhauser (1979) e, più precisamente, Genest e Zidek (1986) hanno dimostrato che nessuna funzione di aggregazione bayesiana soddisfa simultaneamente un insieme di proprietà minimali ragionevoli — un risultato che nella letteratura va sotto il nome di "problema dell'opinione del comitato bayesiano". Il risultato parallelo a quello di Arrow per il contesto bayesiano è che le funzioni di pooling lineare:

$$\Pi^0_C(\theta) = \sum_i w_i \cdot \pi^0_i(\theta), \text{ con } w_i \geq 0 \text{ e } \sum_i w_i = 1$$

soddisfano la proprietà di rispetto dell'unanimità — se tutti gli agenti assegnano probabilità zero a un evento, il meta-modello collettivo fa altrettanto — ma violano la proprietà di Pareto per le scommesse: un portafoglio di scommesse che tutti gli agenti giudicano favorevole può essere sfavorevole sotto la distribuzione aggregata lineare.

Viceversa, le funzioni di pooling geometrico:

$$\Pi^0_C(\theta) \propto \prod_i \pi^0_i(\theta)^{w_i}$$

soddisfano la proprietà di Pareto per le scommesse ma violano la coerenza bayesiana: l'aggiornamento con evidenza comune non commuta con l'aggregazione.

Il risultato di impossibilità dice che non esiste una funzione di aggregazione dei meta-modelli individuali che soddisfi simultaneamente tutte le proprietà desiderabili. Questa è l'istanza del teorema di Lawvere per la metacognizione collettiva: la diagonalizzazione che produce l'impossibilità è la funzione di aggiornamento bayesiano applicata all'aggregazione, che genera il punto fisso impossibile dell'aggregazione coerente e paretiana simultaneamente.

V.4 La struttura delle comunità epistemiche come risposta strutturale all'impossibilità

Dati i risultati di impossibilità per la metacognizione collettiva, le strutture che le comunità epistemiche reali adottano possono essere interpretate come soluzioni approssimate al problema dell'aggregazione — soluzioni che sacrificano consapevolmente alcune delle proprietà desiderabili per ottenere coerenza pratica.

La struttura gerarchica delle comunità epistemiche — con esperti di dominio, revisori inter pares, meta-esperti e policy makers — è una struttura di aggregazione che sacrifica la proprietà di Pareto bayesiano per mantenere la coerenza operativa. In questa struttura, il meta-modello collettivo è determinato prevalentemente dal meta-modello degli esperti di livello più alto, con meccanismi di aggiornamento che incorporano gli input degli agenti di livello inferiore in modo pesato e filtrato. È una struttura di aggregazione che si avvicina alla funzione dittatoriale del teorema di Arrow — con il "dittatore" distribuito tra un insieme di agenti di alta competenza riconosciuta — ma che non è completamente dittatoriale perché i meccanismi di revisione e contestazione permettono l'aggiornamento del meta-modello collettivo.

La struttura del mercato delle credenze — il meccanismo attraverso cui le credenze si propagano e si stabilizzano in reti di agenti che si influenzano reciprocamente — è una soluzione alternativa che sacrifica la razionalità bayesiana per ottenere la robustezza: le credenze che si stabilizzano non sono necessariamente quelle bayesianamente ottimali, ma sono quelle che superano il processo di selezione competitiva nella rete di influenza. Questa struttura è soggetta ai bias noti dell'informazione cascade — la cascata informazionale in cui agenti razionali ignorano le proprie informazioni private in favore dell'osservazione del comportamento dei precedenti — che corrisponde formalmente alla convergenza del sistema verso un punto fisso sub-ottimale determinato dalla struttura della rete anziché dalla qualità delle informazioni individuali.

V.5 Il ruolo della metacognizione individuale nella metacognizione collettiva

La connessione tra la metacognizione individuale e quella collettiva è bidirezionale in modo analogo alla connessione substrato-linguaggio della seconda monografia. La metacognizione individuale — la qualità del meta-modello π^0_i del singolo agente —

determina la qualità del contributo dell'agente al meta-modello collettivo. Ma il meta-modello collettivo a sua volta influenza la metacognizione individuale: le strutture di credenza collettive — i frame culturali, i vocabolari disciplinari, le norme epistemiche di una comunità — orientano la geometria del prior metacognitivo individuale attraverso il meccanismo del termine g_{sub} nell'equazione del sistema accoppiato della seconda monografia.

Si può formalizzare questo come un sistema dinamico accoppiato a tre livelli:

$$\begin{aligned}\dot{m}_{\text{sub}}(t) &= f(m_{\text{sub}}, m_{\text{lang}}, m_{\text{meta}}, e(t)) \\ \dot{m}_{\text{lang}}(t) &= g(m_{\text{sub}}, m_{\text{lang}}, m_{\text{meta}}, c(t)) \\ \dot{m}_{\text{meta}}(t) &= h(m_{\text{sub}}, m_{\text{lang}}, m_{\text{meta}}, \Pi_C(t))\end{aligned}$$

dove $\Pi_C(t)$ è il meta-modello collettivo al tempo t che entra come input nella dinamica del meta-modello individuale. Il punto fisso di questo sistema a tre livelli è la struttura cognitiva stabile dell'agente immerso in una comunità epistemica: la configurazione in cui il substrato, la competenza linguistica e il meta-modello individuale si sono co-adattati reciprocamente e si sono allineati con le strutture collettive della propria comunità.

Questo allineamento con le strutture collettive è il meccanismo della socializzazione cognitiva: un agente che cresce in una certa comunità epistemica tende a sviluppare un meta-modello individuale coerente con il meta-modello collettivo di quella comunità — con la stessa distribuzione delle "certezze" e delle "aree di incertezza", con le stesse strutture di autorità epistemica, con le stesse abitudini di allocazione dell'attenzione metacognitiva. Questa coerenza non è patologica: è il meccanismo attraverso cui la conoscenza si accumula e si trasmette attraverso le generazioni. Ma ha un costo: il meta-modello individuale che si allinea al collettivo perde la capacità di rappresentare le strutture del proprio substrato che non hanno rappresentazione nel meta-modello collettivo — il costo della socializzazione cognitiva è un restringimento del prior metacognitivo individuale nelle direzioni non validate dal collettivo.

Capitolo VI — La logica della dimostrabilità GL e la geometria dell'auto-conoscenza formale

VI.1 Il limite del teorema di Löb come strumento isolato

Il Capitolo II ha impiegato il teorema di Löb come uno dei nodi della tassonomia unificata sotto lo schema di Lawvere: un risultato che pone un limite all'auto-fiducia epistemica del sistema formale, impedendo che l'auto-validazione produca conoscenza nuova. Quella trattazione era corretta ma incompleta in un senso preciso: il teorema di Löb non è un risultato isolato che si può invocare puntualmente — è l'assioma fondante di un intero sistema logico modale, la logica GL (Gödel-Löb), che è il

sistema completo per la logica della dimostrabilità aritmetica. Trattare Löb senza GL è come citare il teorema di Pitagora senza sviluppare la geometria euclidea: si perde la struttura che rende il risultato potente e applicabile.

La logica GL non è un accorgimento tecnico per logici professionisti: è la struttura formale della riflessività cognitiva. Ogni sistema che ragiona sulla propria capacità di ragionare — ogni metacognizione formalizzabile — opera in un ambiente logico la cui geometria è esattamente quella dei frame di Kripke per GL. Sviluppare GL in dettaglio non è quindi un esercizio di completezza enciclopedica: è identificare la geometria propria dello spazio in cui la metacognizione si svolge.

VI.2 Il sistema GL: assiomi e regole

Il sistema modale GL (talvolta denominato G, o L in onore di Löb) è un sistema di logica proposizionale modale. Il suo vocabolario include le variabili proposizionali $p, q, r, \dots$, i connettivi classici $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$, e l'operatore modale unario $\Box$ letto "è dimostrabile che". L'operatore duale $\Diamond\varphi$ è definito come $\neg\Box\neg\varphi$ — "è consistente con il sistema che φ ".

Gli assiomi di GL sono:

$$(K) \quad \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$$

$$(4) \quad \Box\varphi \rightarrow \Box\Box\varphi$$

$$(W) \quad \Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$$

La regola di inferenza aggiuntiva è la Necessitazione:

$$\text{se } \vdash \varphi \text{ allora } \vdash \Box\varphi$$

Il sistema classico K è costituito da (K) più Necessitazione. Il sistema K4 aggiunge (4). Il sistema GL aggiunge (W) a K4.

L'assioma (K) esprime la distribuzione della dimostrabilità sull'implicazione: se è dimostrabile che φ implica ψ , e φ è dimostrabile, allora ψ è dimostrabile. L'assioma (4) esprime la predicabilità della dimostrabilità: se φ è dimostrabile, è dimostrabile che φ è dimostrabile. L'assioma (W) è l'assioma di Löb nella forma modale: se è dimostrabile che "la dimostrabilità di φ implica φ ", allora φ è dimostrabile.

Si osservi che l'assioma di riflessività dei sistemi modali normali più comuni — T: $\Box\varphi \rightarrow \varphi$ — è deliberatamente assente da GL. La sua assenza è il contenuto del secondo teorema di incompletezza di Gödel: il sistema non può dimostrare la propria coerenza, il che significa precisamente che $\Box\perp \rightarrow \perp$ (se è dimostrabile la contraddizione allora

la contraddizione vale, cioè il sistema è inconsistente) non è derivabile in PA. In GL si ha invece il teorema:

$$GL \vdash \neg \Box \neg \Box \perp$$

che si legge: "non è dimostrabile che non sia dimostrabile la falsità" — non è dimostrabile la coerenza del sistema.

VI.3 La semantica di Kripke per GL: frame transitivi e noetheriani

La semantica di Kripke per le logiche modali associa a ogni sistema un insieme di frame. Un frame per una logica modale è una coppia (W, R) dove W è un insieme di "mondi possibili" e $R \subseteq W \times W$ è la relazione di accessibilità. Una valutazione V assegna a ogni variabile proposizionale un sottoinsieme di W (l'insieme dei mondi in cui quella variabile è vera). La verità di una formula modale in un mondo w sotto una valutazione V è definita ricorsivamente, con:

$$(W, R, V), w \models \Box \varphi \Leftrightarrow \text{per ogni } v \in W \text{ con } wRv, (W, R, V), v \models \varphi$$

I frame validi per GL — quelli in cui tutti gli assiomi di GL sono validi — sono caratterizzati da due proprietà:

La *transitività*: se wRv e vRu allora wRu . Corrisponde all'assioma (4): la dimostrabilità è chiusa per composizione. Se da w si "raggiunge" v (v è accessibile da w , cioè φ è dimostrabile in v secondo w) e da v si raggiunge u , allora u è raggiungibile da w .

La *proprietà noetheriana* (o fondatezza convers): non esistono catene infinite ascendenti $w_0Rw_1Rw_2R\ldots$. La relazione R non ha cicli e non ha catene infinitamente ascendenti. Questa è la proprietà che corrisponde all'assioma (W): l'accessibilità è ben fondata in direzione inversa.

Un frame (W, R) transitivo e noetheriano è un frame GL. La semantica di Kripke per GL è *completa e corretta* rispetto alla classe dei frame GL: una formula φ è derivabile in GL se e soltanto se è valida in tutti i frame GL.

La struttura geometrica dei frame GL è quella di un ordine parziale strict (irreflesso, transitivo) privo di catene infinite ascendenti — essenzialmente un ordine parziale ben fondato in direzione inversa. I mondi si organizzano in "livelli" dove ogni mondo può accedere solo a mondi di livello superiore, senza la possibilità di risalire indefinitamente. Questa è la geometria dello spazio della riflessività metacognitiva: ogni atto di auto-riflessione porta a un mondo più alto nella gerarchia, e la noetherianità garantisce che la riflessività non sia circolare.

VI.4 Il teorema di completezza aritmetica di Solovay

Il teorema di Solovay (1976) è il risultato che trasforma GL da sistema logico formale in strumento di analisi metacognitiva con interpretazione aritmetica precisa. Il teorema stabilisce che GL cattura esattamente il comportamento del predicato di dimostrabilità di $\text{Bew}(x)$ nell'aritmetica di Peano.

Si definisce un'interpretazione aritmetica $*$ come una funzione $*$ che:

- Associa a ogni variabile proposizionale p_i una sentenza aritmetica p_i^* della PA
- Si estende ai connettivi classici in modo composizionale: $(\varphi \wedge \psi)^* = \varphi^* \wedge \psi^*$, ecc.
- Traduce l'operatore $\Box$ con il predicato di dimostrabilità: $(\Box\varphi)^* = \text{Bew}(\ulcorner \varphi^* \urcorner)$

Teorema di Solovay (1976): Per ogni formula modale φ , sono equivalenti:

- (i) $\text{GL} \vdash \varphi$
- (ii) Per ogni interpretazione aritmetica $*$, $\text{PA} \vdash \varphi^*$

Il teorema ha due direzioni. La *correttezza* (i $\Rightarrow$ ii) afferma che ogni teorema di GL diventa un teorema di PA sotto qualsiasi interpretazione aritmetica: ogni proprietà formale della dimostrabilità catturata da GL è effettivamente vera di Bew in PA. La *completezza* (ii $\Rightarrow$ i) afferma che GL cattura *tutti* i teoremi modali di PA su Bew : se una formula modale φ non è derivabile in GL, esiste un'interpretazione aritmetica sotto cui φ^* non è dimostrabile in PA.

La completezza è la direzione profonda: essa afferma che GL è *massimalmente informativo* sulla dimostrabilità in PA. Nessun sistema modale strettamente più forte di GL è valido aritmeticamente: ogni assioma aggiunto a GL produce teoremi che non sono validi per Bew in PA. In particolare, l'assioma T ($\Box\varphi \rightarrow \varphi$, la riflessività) non è valido aritmeticamente — coerentemente con il secondo teorema di incompletezza di Gödel.

Il corollario per la metacognizione è il seguente. La struttura formale di qualsiasi sistema che ragiona sulla propria capacità dimostrativa — o, per estensione, sulla propria struttura cognitiva accessibile all'auto-osservazione — è la struttura di un frame GL. La geometria dell'auto-conoscenza formalizzabile è la geometria degli ordini parziali strict ben fondati. Ogni tentativo di estendere quella geometria aggiungendo riflessività (il sistema può auto-validarsi completamente) produce inconsistenza; ogni tentativo di aggiungere catene infinite di auto-riflessione produce sistemi non fondati.

VI.5 Le conseguenze di Solovay per la torre bayesiana gerarchica

Il teorema di Solovay illumina la struttura della torre bayesiana gerarchica introdotta nel Capitolo I in modo che non era anticipabile senza la logica GL. La torre dei livelli $\pi^0, \pi^1, \pi^2, \dots$ non è semplicemente una successione gerarchica di distribuzioni di

probabilità: è la struttura semantica di una sequenza di mondi accessibili in un frame GL.

Il livello π^0 è il mondo w_0 — il mondo base in cui il sistema ha credenze sui propri parametri cognitivi θ . Il livello π^1 è il mondo w_1 accessibile da w_0 — il mondo in cui il sistema ha credenze sulla bontà delle proprie credenze di livello base. La relazione di accessibilità R tra i livelli è transitiva e noetheriana per costruzione: la torre è ben fondata verso il basso (ogni livello è fondato sul livello inferiore) e ha profondità finita (non ci sono catene infinite di auto-riflessione).

Formalmente, si definisce il *frame metacognitivo* (W_K, R_K) associato a un sistema con profondità metacognitiva K come:

$$W_K = \{w_0, w_1, \dots, w_K\}$$

$$w_i R_K w_j \Leftrightarrow i < j$$

Questo è chiaramente un frame GL: la relazione è l'ordine strict su $\{0, \dots, K\}$, che è transitivo e noetheriano. La dimostrabilità in questo frame è la derivabilità nel sistema metacognitivo al livello corrispondente.

Il teorema di Solovay ha ora una lettura metacognitiva precisa. $GL \vdash \varphi$ significa che la proprietà φ dell'auto-conoscenza è valida in *tutti* i frame GL — valida per qualsiasi profondità metacognitiva K , indipendentemente dalla struttura specifica del sistema. In aritmetica: la proprietà è una verità necessaria su Bew in PA. La completezza di Solovay afferma che GL è massimalmente forte in questo senso: non ci sono verità necessarie sull'auto-conoscenza formalizzabile che GL non catturi.

VI.6 Il secondo teorema di incompletezza nella geometria GL

Il secondo teorema di incompletezza di Gödel — che PA non può dimostrare la propria coerenza, cioè $PA \not\vdash \neg \text{Bew}(\ulcorner \bot \urcorner)$ — ha una formulazione modale elegante in GL:

$$GL \not\vdash \neg \Box \bot$$

La formula $\neg \Box \bot$ afferma "è falso che sia dimostrabile una contraddizione", ovvero "il sistema è coerente". Questa formula non è un teorema di GL, il che — per il teorema di Solovay — significa esattamente che PA non può dimostrare la propria coerenza sotto nessuna interpretazione aritmetica: per ogni scelta di p_i^* , la traduzione di $\neg \Box \bot$ in PA non è sempre dimostrabile.

Nei frame GL, questo corrisponde al fatto che $\neg \Box \bot$ non è valida in ogni frame GL: esiste un frame GL e un mondo in esso in cui $\neg \Box \bot$ è falsa. Costruzione esplicita: si prenda il frame a un solo mondo $W = \{w_0\}$ con relazione vuota $R = \emptyset$. In w_0 , $\Box \bot$ è

vacuamente vera (non esistono mondi accessibili da w_0 , quindi $\neg \Box \perp$ richiede un mondo accessibile in cui $\perp$ sia vera — condizione soddisfatta vacuamente in senso rovesciato: $\neg \Box \perp$ equivale a $\Diamond \top$, che richiede un mondo accessibile, e non ce ne sono). Dunque $\Box \perp$ è vera e $\neg \Box \perp$ è falsa in w_0 .

La struttura del frame a mondo singolo con relazione vuota è il modello della "self-referential incoherence": il sistema non ha accesso a nessun mondo metacognitivo superiore — non ha livelli di auto-riflessione — e in questo sistema il predicato di coerenza fallisce. È la geometria dell'assenza di metacognizione: un sistema che non può raggiungere nessun punto di vista esterno su se stesso non può nemmeno formulare la propria coerenza come proprietà verificabile.

VI.7 La logica GL e i gradi della riflessività metacognitiva

La logica GL permette di classificare le proprietà metacognitive in base alla loro "profondità modale" — il numero di operatori $\Box$ annidati necessari per esprimerle. Le formule con un solo $\Box$ esprimono proprietà di primo ordine della dimostrabilità; le formule con $\Box\Box$ esprimono proprietà di secondo ordine; e così via.

La gerarchia delle proprietà metacognitive per profondità modale è:

Profondità 0: le proposizioni non modali — le credenze del sistema su se stesso esprimibili senza riferimento alla propria capacità dimostrativa. Corrispondono ai parametri cognitivi θ nel framework bayesiano.

Profondità 1 ($\Box\varphi$): proprietà della prima auto-riflessione — ciò che il sistema può dimostrare di se stesso in un singolo passo di auto-osservazione. Corrisponde al meta-prior $\pi^0(\theta)$.

Profondità 2 ($\Box\Box\varphi$): proprietà della seconda auto-riflessione — ciò che il sistema può dimostrare sulla propria capacità di dimostrare. Corrisponde all'iper-prior $\pi^1(\pi^0)$.

Profondità k ($\Box^k\varphi$): il k -esimo livello della torre metacognitiva.

Un risultato cruciale di GL è che la gerarchia per profondità modale è *strettamente crescente*: per ogni k , esistono formule di profondità $k+1$ che non sono equivalenti in GL a nessuna formula di profondità k . La gerarchia metacognitiva non collassa: i livelli superiori della torre esprimono genuinamente più di quelli inferiori, e non esiste una "formula sommaria" finita che catturi tutte le proprietà metacognitive di tutti i livelli.

Questo risultato ha la seguente lettura per la tower bayesiana: non esiste una distribuzione di probabilità finita-dimensionale che catturi, senza perdita, l'intera struttura della torre Π^∞ . Ogni troncatura a profondità K perde genuine proprietà metacognitive che si trovano ai livelli superiori. La perdita non è tecnica ma strutturale: è la perdita di proprietà reali dell'auto-conoscenza che sono vere del sistema ma inesprimibili al livello K .

Capitolo VII — Analisi ordinale e misura della portata metacognitiva

VII.1 L'ordinale proof-teorico come misura della forza riflessiva

Il secondo teorema di incompletezza di Gödel stabilisce che nessun sistema formale coerente e sufficientemente espressivo può dimostrare la propria coerenza. Questo stabilisce una gerarchia tra i sistemi: un sistema S_2 più forte di S_1 può dimostrare la coerenza di S_1 , ma non la propria. La gerarchia della forza proof-teorica è misurata dagli *ordinali proof-teorici*.

L'ordinale proof-teorico di un sistema S , indicato con $|S|$, è il più piccolo ordinale che S non può dimostrare essere ben fondato. Più precisamente: S può dimostrare la ben-fondatezza di qualsiasi ordinale $\alpha < |S|$ mediante induzione transfinita fino ad α , ma non può dimostrare la ben-fondatezza di $|S|$ stesso senza superare la propria forza proof-teorica.

L'ordinale proof-teorico è la misura quantitativa della "portata metacognitiva" del sistema: quanto in alto nella gerarchia degli ordinali il sistema può "vedere". Un sistema con ordinale proof-teorico α può certificare la consistenza di tutti i sistemi con ordinale inferiore ma non la propria. La metacognizione formale ha una misura cardinale precisa: non è un tutto-o-nulla ma una quantità ordinale.

VII.2 ε_0 e il teorema di Gentzen per l'aritmetica di Peano

Il risultato fondamentale dell'analisi ordinale per PA è il teorema di Gentzen (1936):

Teorema di Gentzen: Il sistema PA è consistente se e soltanto se il principio di induzione transfinita fino a ε_0 è valido, dove ε_0 è il primo ordinale ε tale che $\omega^\varepsilon = \varepsilon$.

L'ordinale ε_0 è definito come il limite della sequenza:

$$\omega^0 = 1, \omega^1 = \omega, \omega^2 = \omega^\omega, \omega^3 = \omega^{\omega^\omega}, \dots$$

$$\varepsilon_0 = \sup_{n < \omega} \omega^{\uparrow n} = \omega^{\{\omega^{\{\omega^{\{\dots\}}}\}} \text{ (torre infinita)}$$

In notazione di Cantor, ε_0 è il più piccolo punto fisso della mappa $\alpha \mapsto \omega^\alpha$. È un ordinale numerabile — ha cardinalità $\aleph_0$ — ma la sua struttura interna è di complessità straordinaria: ogni ordinale $\alpha < \varepsilon_0$ ammette una rappresentazione in forma normale di Cantor nella base ω con esponenti strettamente minori di α .

Il teorema di Gentzen afferma dunque:

$$|PA| = \varepsilon_0$$

PA può dimostrare la ben-fondatezza di ogni ordinale $\alpha < \varepsilon_0$, ma non di ε_0 stesso: la consistenza di PA equivale all'induzione fino a ε_0 , e questa induzione non è dimostrabile in PA (che sarebbe la propria consistenza). In PRA (aritmetica ricorsiva primitiva), sistema più debole di PA, si può dimostrare la consistenza di PA aggiungendo esattamente il principio di induzione transfinita $TI(\varepsilon_0)$: la forza aggiuntiva necessaria è misurata con precisione ordinale.

VII.3 La gerarchia degli ordinali proof-teorici come mappa della metacognizione

La tabella degli ordinali proof-teorici dei principali sistemi fornisce una mappa quantitativa della gerarchia dell'auto-conoscenza formale:

PRA (aritmetica ricorsiva primitiva): $|PRA| = \omega^\omega$

PA (aritmetica del primo ordine): $|PA| = \varepsilon_0$

ACA_0 (aritmetica del secondo ordine, base): $|ACA_0| = \varepsilon_0$ (stessa di PA, sorprendentemente)

ATR_0 (ricorsione transfinita): $|ATR_0| = \Gamma_0$

$\Pi^1_1\text{-}CA_0$ (comprensione Π^1_1): $|\Pi^1_1\text{-}CA_0| = \psi(\Omega^\Omega)$

PA + "PA è consistente": $|PA + \text{Con}(PA)| = \varepsilon_1 = \varepsilon_{-\{ \varepsilon_0 \}}$

dove Γ_0 è l'ordinale di Feferman-Schütte (il primo ordinale non raggiungibile per via predicativa), e $\psi(\Omega^\Omega)$ è un ordinale ben più grande nella gerarchia delle funzioni di collasso ordinale.

La sequenza PA, PA + Con(PA), PA + Con(PA + Con(PA)), ... genera ordinali:

$\varepsilon_0, \varepsilon_1, \varepsilon_2, \dots \rightarrow \varepsilon_{-\{ \varepsilon_0 \}}, \dots \rightarrow \varepsilon_{-\{ \varepsilon_{-\{ \varepsilon_0 \}} \}}, \dots$

che corrispondono esattamente alla torre metacognitiva: ogni livello della torre — ogni sistema che dimostra la consistenza del livello inferiore — ha ordinale proof-teorico strettamente maggiore. La torre metacognitiva bayesiana $\pi^0, \pi^1, \pi^2, \dots$ ha un'immagine ordinale precisa $\varepsilon_0, \varepsilon_1, \varepsilon_2, \dots$ nella gerarchia degli epsilon-numeri di Cantor.

VII.4 La connessione tra profondità bayesiana K e ordinale proof-teorico

Si stabilisce ora la connessione precisa tra la profondità metacognitiva K della torre bayesiana del Capitolo I e la struttura ordinale del Capitolo VII.

Un sistema cognitivo con profondità metacognitiva K — che mantiene rappresentazioni esplicite fino al livello π^K — corrisponde, in termini proof-teorici, a un sistema formale di forza ordinale $|S_K|$ dove:

$|S_0| = \varepsilon_0$ (il sistema ha self-knowledge di base, analogo a PA)

$|S_k| = \varepsilon_k$ (il sistema può certificare la consistenza di tutti i livelli inferiori)

La profondità metacognitiva K determina fino a quale ordinale ε_K il sistema può "vedere" la propria struttura. La *portata metacognitiva* del sistema è l'ordinale $|S_K|$: la misura di quanta struttura dell'auto-conoscenza il sistema è in grado di certificare.

La conseguenza fondamentale è che la torre bayesiana gerarchica non è uno strumento neutro di rappresentazione delle credenze: è un dispositivo che genera, a ogni livello, un sistema di forza proof-teorica strettamente maggiore. Aggiungere un livello alla torre — passare da K a $K+1$ — non è un'operazione banale di arricchimento della rappresentazione: è il passaggio a un sistema formale più forte, con una nuova "assioma di consistenza" che il livello inferiore non poteva dimostrare.

VII.5 Il teorema di Friedman-Rathjen e la barriera predicativa

La barriera di Γ_0 — l'ordinale di Feferman-Schütte — ha un significato metacognitivo speciale. Γ_0 è il primo ordinale che non è raggiungibile attraverso operazioni predicative: non può essere definito senza circolarità, senza usare nella propria definizione ordinali che presuppongono già ordinali di dimensione comparabile.

Il programma predicativista di Weyl e Feferman identifica Γ_0 come il limite superiore della matematica giustificabile senza presupporre l'esistenza di oggetti impredicativi. Nel contesto metacognitivo, la barriera predicativa corrisponde al limite della metacognizione "fondata": un sistema la cui auto-conoscenza rimane al di sotto di Γ_0 può giustificare la propria struttura senza circolarità; un sistema che supera Γ_0 inevitabilmente usa la propria struttura per giustificare se stesso.

I sistemi cognitivi umani — nella misura in cui hanno una struttura metacognitiva formalizzabile — operano presumibilmente al di sotto della barriera predicativa per quanto riguarda la metacognizione esplicita accessibile alla coscienza. La metacognizione implicita del substrato subpropositivo, operante in $\mathbb{R}^n$, non è soggetta a questi limiti formali perché non è un sistema dimostrativo: è un sistema di rappresentazione distribuita che non genera derivazioni. La barriera di Γ_0 si applica alla componente proposizionale della metacognizione — al ragionamento esplicito su se stessi — non al substrato.

VII.6 Ordinali, apprendimento e la dinamica della portata metacognitiva

La portata metacognitiva — misurata dall'ordinale proof-teorico $|S_K|$ — non è fissa nel corso della vita di un agente: si espande con l'apprendimento. Ogni volta che il sistema acquisisce la capacità di ragionare su un livello superiore della propria struttura — ogni volta che la profondità metacognitiva K aumenta — l'ordinale proof-teorico associato sale di un gradino nella gerarchia degli epsilon-numeri.

L'equazione dinamica per l'ordinale metacognitivo è:

$$|S_K(t+1)| = \epsilon_{\{ |S_K(t)| \}} \quad (\text{aumento di un livello riflessivo})$$

dove ϵ_α denota il α -esimo epsilon-numero. Questa è la mappa di Cantor che porta un ordinale al successivo punto fisso di $\alpha \mapsto \omega^\alpha$ sopra di esso: la transizione da ϵ_n a ϵ_{n+1} corrisponde esattamente all'acquisizione della capacità di dimostrare la consistenza del sistema di livello n .

La velocità a cui un sistema cognitivo può espandere la propria portata metacognitiva è limitata dal costo computazionale dell'aggiornamento bayesiano al livello superiore — la variabile di co-stato v del problema di controllo ottimale del Capitolo IV misura esattamente questo costo — e dalla plasticità sinaptica disponibile per la modifica della struttura metacognitiva. Il punto di equilibrio della portata metacognitiva — il valore K^* a cui il sistema converge — è il punto fisso della funzione di utilità netta dell'approfondimento metacognitivo, determinato dal bilanciamento tra il valore aggiunto di un livello superiore e il suo costo computazionale.

Capitolo VIII — Il lemma diagonale di Gödel-Carnap: derivazione e applicazioni

VIII.1 Il lemma diagonale come strumento chirurgico

Nel Capitolo II, il teorema di punto fisso di Lawvere è stato presentato come lo schema categoriale che unifica tutti i risultati di impossibilità per l'auto-conoscenza. Questa unificazione è genuina e potente, ma opera a un livello di astrazione che nasconde la struttura costruttiva dei singoli risultati. Il lemma diagonale di Gödel-Carnap è lo strumento che porta quella struttura costruttiva alla luce: esso produce, per ogni proprietà espressa nel linguaggio del sistema, una *sentenza che afferma di se stessa di avere quella proprietà*. È il meccanismo attraverso cui l'auto-riferimento viene costruito concretamente nell'aritmetica, e la sua derivazione esplicita è necessaria per capire perché ogni tentativo di auto-rappresentazione completa produce l'oggetto impossibile.

VIII.2 La funzione di sostituzione e la sua rappresentabilità

Il lemma diagonale si basa sulla rappresentabilità di una funzione aritmetica specifica: la funzione di sostituzione. Si opera nell'aritmetica di Peano con una codifica di Gödel fissata: ogni formula φ del linguaggio di PA ha un codice di Gödel $\ulcorner \varphi \urcorner \in \mathbb{N}$, e ogni numero naturale n ha un termine numerale corrispondente $\bar{n}$ nel linguaggio di PA.

Si definisce la funzione di sostituzione:

$$\text{sub} : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$$

$$\text{sub}(\ulcorner \varphi(x) \urcorner, n) = \ulcorner \varphi(\bar{n}) \urcorner$$

la funzione che, dati il codice di una formula $\varphi(x)$ con la variabile libera x e un numero n , produce il codice della formula ottenuta sostituendo $\bar{n}$ per x in φ . Questa funzione è *primitiva ricorsiva* — si calcola attraverso una procedura di riscrittura sintattica che opera sulla struttura del codice di Gödel in modo algebrico elementare — e quindi è *rappresentabile* in PA: esiste una formula $\text{Sub}(x, y, z)$ del linguaggio di PA tale che:

$$\text{per ogni } n, m \in \mathbb{N} : \text{sub}(n, m) = k \Leftrightarrow \text{PA} \vdash \text{Sub}(\bar{n}, \bar{m}, \bar{k})$$

La rappresentabilità di sub è il punto tecnico su cui regge tutto il lemma: il sistema può parlare delle proprie formule perché può calcolare, al proprio interno, la funzione che sostituisce numeri nelle formule.

VIII.3 Derivazione del lemma diagonale

Lemma diagonale (Gödel-Carnap): Per ogni formula $\psi(x)$ del linguaggio di PA con esattamente una variabile libera x , esiste una sentenza φ (senza variabili libere) tale che:

$$\text{PA} \vdash \varphi \leftrightarrow \psi(\ulcorner \varphi \urcorner)$$

La sentenza φ *afferma di se stessa di avere la proprietà espressa da ψ* .

Dimostrazione:

Passo 1 — Costruzione della formula diagonale. Sia $\psi(x)$ la formula data. Si definisce la formula:

$$\theta(x) := \psi(\text{Sub}(x, x, \cdot))$$

dove $\text{Sub}(x, x, \cdot)$ è la formula che esprime "il risultato della sostituzione di $\bar{x}$ per x nella formula di codice x ", ovvero $\text{Sub}(x, x, z)$ rappresenta $\text{sub}(n, n)$ per $n = \text{val}(x)$. Più precisamente, si scrive:

$$\theta(x) := \exists z (\text{Sub}(x, x, z) \wedge \psi(z))$$

La formula $\theta(x)$ dice: "esiste z che è il risultato della sostituzione del numero x nella formula di codice x , e z soddisfa ψ ".

Passo 2 — Applicazione alla propria codifica. Sia $n_0 = \ulcorner \theta(x) \urcorner$ il codice di Gödel della formula $\theta(x)$. Si definisce:

$$\varphi := \theta(n_0) = \exists z (\text{Sub}(n_0, n_0, z) \wedge \psi(z))$$

Passo 3 — Calcolo del punto fisso. Si calcola il valore di $\text{sub}(n_0, n_0)$:

$$\text{sub}(n_0, n_0) = \text{sub}(\ulcorner \theta(x) \urcorner, n_0) = \ulcorner \theta(\bar{n}_0) \urcorner = \ulcorner \varphi \urcorner$$

Poiché sub è rappresentata da Sub , PA dimostra:

$$\text{PA} \vdash \text{Sub}(n_0, n_0, \ulcorner \varphi \urcorner) \wedge \forall z (\text{Sub}(n_0, n_0, z) \rightarrow z = \ulcorner \varphi \urcorner)$$

Passo 4 — Derivazione dell'equivalenza. Da $\varphi = \theta(n_0) = \exists z (\text{Sub}(n_0, n_0, z) \wedge \psi(z))$ e dal fatto che $\text{Sub}(n_0, n_0, z)$ implica $z = \ulcorner \varphi \urcorner$, si ha:

$$\text{PA} \vdash \varphi \leftrightarrow \psi(\ulcorner \varphi \urcorner)$$

Questo completa la dimostrazione. $\square$

La costruzione è esattamente l'argomento diagonale di Cantor applicato alle formule: φ è la formula che "si indica da sola" nella formula ψ . Il numero n_0 è il numero che codifica la formula che produce φ quando viene auto-applicata: è l'analogo aritmetico del quine — il programma che stampa il proprio codice sorgente.

VIII.4 La formula di Gödel come applicazione del lemma diagonale

La formula di Gödel si ottiene applicando il lemma diagonale a $\psi(x) = \neg \text{Bew}(x)$, dove $\text{Bew}(x)$ è il predicato di dimostrabilità (x è il codice di una formula dimostrabile in PA).

Il lemma diagonale produce una sentenza G tale che:

$$PA \vdash G \leftrightarrow \neg \text{Bew}(\ulcorner G \urcorner)$$

G afferma di se stessa: "non sono dimostrabile in PA". La formula di Gödel è precisamente il punto fisso di $\neg \text{Bew}$, ed è costruita con la procedura esplicita del lemma.

Da questa equivalenza e dall'ipotesi che PA sia coerente, si derivano entrambi i teoremi di incompletezza:

Primo teorema: G non è dimostrabile in PA. Supposto che $PA \vdash G$, allora $PA \vdash \text{Bew}(\ulcorner G \urcorner)$ per la definizione di Bew . Ma $PA \vdash G \leftrightarrow \neg \text{Bew}(\ulcorner G \urcorner)$ implica $PA \vdash \neg \text{Bew}(\ulcorner G \urcorner)$, contraddizione con la coerenza di PA. Quindi $PA \not\vdash G$. Ma allora $\neg \text{Bew}(\ulcorner G \urcorner)$ è vera (aritmeticamente), e per l'equivalenza G è vera — vera ma non dimostrabile.

Secondo teorema: $PA \not\vdash \text{Con}(PA)$. La coerenza di PA è $\neg \text{Bew}(\ulcorner \perp \urcorner)$. Si può mostrare che $PA \vdash \text{Con}(PA) \rightarrow G$: se PA fosse coerente (che PA può formalizzare come $\neg \text{Bew}(\ulcorner \perp \urcorner)$), allora G sarebbe vera. Ma $PA \not\vdash G$, quindi $PA \not\vdash \text{Con}(PA)$.

VIII.5 Il predicato di verità di Tarski come applicazione del lemma diagonale

Il teorema di indefinibilità della verità di Tarski si ottiene applicando il lemma a $\psi(x) = \neg \text{True}(x)$, dove si suppone per assurdo che $\text{True}(x)$ sia un predicato del linguaggio di PA che soddisfa lo schema di Tarski $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$.

Il lemma diagonale produce una sentenza λ tale che:

$$PA \vdash \lambda \leftrightarrow \neg \text{True}(\ulcorner \lambda \urcorner)$$

La sentenza λ afferma di se stessa "non sono vera" — il paradosso del Mentitore formalizzato. Dallo schema di Tarski:

$$PA \vdash \text{True}(\ulcorner \lambda \urcorner) \leftrightarrow \lambda \text{ e } PA \vdash \lambda \leftrightarrow \neg \text{True}(\ulcorner \lambda \urcorner)$$

Si ottiene $PA \vdash \text{True}(\ulcorner \lambda \urcorner) \leftrightarrow \neg \text{True}(\ulcorner \lambda \urcorner)$, che è una contraddizione. Quindi True non può essere definito in PA soddisfacendo lo schema di Tarski: il sistema non può contenere al proprio interno un predicato di verità per le proprie sentenze.

La struttura del paradosso del Mentitore non è un'anomalia retorica: è la costruzione concreta del punto fisso impossibile che il teorema di Lawvere prevede in astratto. Il lemma diagonale è il meccanismo che produce il Mentitore nella forma rigorosa dell'aritmetica.

VIII.6 Il teorema di Löb come punto fisso del lemma diagonale

Il teorema di Löb si dimostra anch'esso attraverso il lemma diagonale, in una versione che mostra esplicitamente come la formula di Löb sia il punto fisso di $\Box\varphi \rightarrow \varphi$.

Si vuole dimostrare: per ogni sentenza φ , se $PA \vdash \Box\varphi \rightarrow \varphi$ allora $PA \vdash \varphi$.

Si applica il lemma diagonale a $\psi(x) = (\text{Bew}(x) \rightarrow \varphi)$, ottenendo una sentenza ℓ tale che:

$$PA \vdash \ell \leftrightarrow (\text{Bew}(\ulcorner \ell \urcorner) \rightarrow \varphi)$$

ℓ è la sentenza che dice "se sono dimostrabile, allora φ è vera" — una sorta di promessa condizionale.

La dimostrazione procede in quattro passi:

Passo 1: $PA \vdash \text{Bew}(\ulcorner \ell \urcorner) \rightarrow \text{Bew}(\ulcorner \text{Bew}(\ulcorner \ell \urcorner) \urcorner)$ per l'assioma (4) della dimostrabilità.

Passo 2: $PA \vdash \text{Bew}(\ulcorner \ell \urcorner) \rightarrow \text{Bew}(\ulcorner \varphi \urcorner)$. Dalla definizione di ℓ , $PA \vdash \ell \leftrightarrow (\text{Bew}(\ulcorner \ell \urcorner) \rightarrow \varphi)$, quindi $PA \vdash \text{Bew}(\ulcorner \ell \leftrightarrow (\text{Bew}(\ulcorner \ell \urcorner) \rightarrow \varphi) \urcorner)$, e la dimostrabilità si propaga lungo l'equivalenza.

Passo 3: Dall'ipotesi $PA \vdash \text{Bew}(\ulcorner \varphi \urcorner) \rightarrow \varphi$ e dal Passo 2, $PA \vdash \text{Bew}(\ulcorner \ell \urcorner) \rightarrow \varphi$.

Passo 4: Dall'equivalenza di ℓ con $(\text{Bew}(\ulcorner \ell \urcorner) \rightarrow \varphi)$ e dal Passo 3, $PA \vdash \ell$. Per il Passo 2 e la riflessione, $PA \vdash \varphi$. $\square$

Il punto fisso ℓ del lemma diagonale per la proprietà $\text{Bew}(x) \rightarrow \varphi$ è il nucleo della dimostrazione del teorema di Löb: è la sentenza che "si impegna" ad essere vera se dimostrabile, e questa struttura di auto-impegno — costruita esplicitamente dalla diagonalizzazione — è ciò che forza la dimostrabilità diretta di φ .

VIII.7 Il lemma diagonale nella semantica dei mondi possibili

La struttura del lemma diagonale può essere trasportata nella semantica dei mondi possibili di Kripke, dove acquista un'interpretazione che connette il risultato tecnico con la metacognizione epistemica degli agenti.

In un frame di Kripke (W, R) con una valutazione V , il lemma diagonale generalizzato afferma: per ogni proprietà $P \subseteq W \times W$ (una relazione tra mondi), esiste un mondo $w^* \in W$ tale che:

$$w^* \in P(w^*)$$

cioè w^* ha la proprietà che descrive la propria relazione con se stesso. Questo è il punto fisso diagonale nella semantica dei mondi possibili.

Per la metacognizione: si interpretino i mondi come stati metacognitivi del sistema (configurazioni di $\pi^0, \pi^1, \dots$) e la relazione R come la relazione di accessibilità del frame GL. La proprietà $P(w) = \{v : v \models \psi(\ulcorner \psi \urcorner)\}$ è la proprietà di soddisfare la formula ψ applicata al codice della formula ψ valida in v . Il lemma diagonale produce uno stato metacognitivo w^* che soddisfa esattamente la proprietà che descrive la propria auto-valutazione: è il formalismo dell'*autoconsapevolezza stabile* — lo stato in cui il sistema ha una rappresentazione di se stesso che è, in senso preciso, il punto fisso della propria funzione di auto-valutazione.

Questo stato non è necessariamente raggiungibile o stabile: è il punto fisso formale, che può cadere fuori dalla regione dei mondi accessibili del frame GL (non è raggiungibile nel frame metacognitivo del sistema) o può essere il punto fisso di una funzione che non ha punti fissi nel dominio rilevante (come la negazione, che produce le formule di Gödel e Tarski). La teoria della metacognizione competente è la teoria di quale punto fisso il sistema mira a raggiungere e quali strutture del frame GL permettono di farlo in modo coerente.

VIII.8 La diagonalizzazione come struttura ricorrente nel ciclo

Il lemma diagonale ha connessioni precise con tutti e quattro i temi del ciclo di monografie che l'accademica ha costruito.

Nella prima monografia — il malinteso come struttura cognitiva — la diagonalizzazione corrisponde al problema dell'auto-riferimento nella comunicazione: quando l'agente A tenta di modellare come l'agente B lo modella (il meta-gioco comunicativo), genera una struttura regressiva la cui struttura formale è quella del lemma diagonale applicato allo spazio degli stati mentali. Il punto fisso della mutua modellazione — il punto in cui A ha un modello di B che è identico al modello che B ha del modello di A — è l'equilibrio del gioco comunicativo, ma la sua raggiungibilità è soggetta agli stessi limiti della diagonalizzazione.

Nella seconda monografia — il substrato subpropositivo — la diagonalizzazione corrisponde alla struttura della generalizzazione geometrica nello spazio di attivazione $\mathbb{R}^n$: la funzione di attivazione del sistema su input che codificano il sistema stesso è il punto fisso del sistema di Brouwer applicato alla mappa di auto-attivazione, con le proprietà di punto fisso garantite dalla compattezza e convessità della varietà cognitiva M .

Nella quarta monografia — l'attenzione come operatore di selezione del gradiente — la diagonalizzazione corrisponde al problema dell'auto-attenzione: quando il meccanismo di attenzione seleziona i propri prior da aggiornare, genera una struttura di auto-riferimento la cui complessità è bounded dal lemma diagonale nel senso che nessun meccanismo di auto-attenzione può selezionare ottimamente i propri prior per tutti i possibili prior iniziali.

La struttura diagonale è il filo rosso formale del ciclo: ogni monografia affronta, nel proprio dominio, la struttura del punto fisso impossibile che emerge quando un sistema sufficientemente espressivo tenta di rappresentare completamente se stesso. Il ciclo è un'esplorazione sistematica delle forme che questa struttura assume — nel linguaggio della teoria dei giochi, degli spazi di attivazione, della logica modale, e dell'ottimizzazione stocastica.

Capitolo IX — Prior di Solomonoff, problema del frame e connessioni con il ciclo

IX.1 Il prior di Solomonoff come limite infattibile dell'ottimalità bayesiana

Nella struttura bayesiana gerarchica del Capitolo I, il prior metacognitivo $\pi^0(\theta)$ è una distribuzione di credenza sui parametri cognitivi del sistema. Una domanda naturale è: quale prior è ottimale dal punto di vista della teoria bayesiana — quale distribuzione assegna probabilità iniziali che producono il migliore aggiornamento a lungo termine per qualsiasi successione di osservazioni?

La risposta teoricamente precisa è il *prior di Solomonoff* (Solomonoff, 1964), definito come:

$$M(x) = \sum_{\{p : U(p) = x\}} 2^{-|p|}$$

dove U è una macchina di Turing universale prefissa fissata, p varia su tutti i programmi per cui U produce l'output x , e $|p|$ è la lunghezza del programma p in bit. $M(x)$ è la misura di probabilità universale: la probabilità che U produca x quando il suo input è una sequenza di bit casuali equiprobabili.

Il *teorema di dominanza* di Solomonoff afferma: per ogni distribuzione di probabilità computable P e per ogni stringa x :

$$M(x) \geq c_P \cdot P(x)$$

dove $c_P > 0$ dipende solo da P . M domina moltiplicativamente ogni prior computable. La connessione con la complessità di Kolmogorov è:

$$-\log_2 M(x) \approx KC(x) \quad (\text{a meno di una costante additiva dipendente da } U)$$

Il prior di Solomonoff assegna probabilità $2^{-KC(x)}$ alla stringa x : è la formalizzazione rigorosa del rasoio di Occam come misura di probabilità.

IX.2 L'infattibilità computazionale del prior di Solomonoff

Il prior di Solomonoff è *non computabile*: $M(x) \geq c \cdot P_{\text{fermata}}(x)$, e il problema della fermata è indecidibile. M è tuttavia *upper semi-computable*: esiste una sequenza di approssimazioni computabili $M_1, M_2, M_3, \dots$ con $M_n(x) \downarrow M(x)$ per ogni x , con convergenza da sopra arbitrariamente lenta.

La distanza dal prior ottimale del miglior prior computabile costruibile con budget di complessità K è:

$$D_{KL}(M \parallel M_K) \geq \Omega(1/K)$$

Più corto è il programma del prior, più grande è la divergenza dall'ottimo.

Questo introduce un vincolo sulla funzione di tasso-distorsione metacognitiva del Capitolo III. La distorsione minima raggiungibile non è solo quella di Shannon: è bounded inferiormente dalla componente algoritmica:

$$D_{\min} \geq \max(D_{\text{Shannon}}, D_{\text{Chaitin}})$$

dove $D_{\text{Chaitin}} = 1 - 2^{-(KC(\theta^*) - K)}$ per un sistema con prior metacognitivo a complessità K . Per sistemi con $K \ll KC(\theta^*)$, questa componente domina. I parametri cognitivi la cui complessità supera K sono assegnati probabilità zero dal prior metacognitivo: non sono nemmeno formulabili come ipotesi.

IX.3 Il prior di Solomonoff e il substrato subpropositivo

La conoscenza procedurale esperta — il pianista, il chirurgo, l'atleta — corrisponde, nello spazio di attivazione $\mathbb{R}^n$ della seconda monografia, a vettori v con alta complessità algoritmica $KC(v)$: la struttura fine delle traiettorie esperte non ammette descrizione breve. Il prior metacognitivo M_K assegna a tali stati probabilità $M_K(v) \approx 0$ se $KC(v) \gg K$. Il sistema esperto non riesce a formulare come ipotesi il proprio stato di expertise procedurale.

Questa è la formalizzazione dell'*auto-opacità radicale*: non la mancanza di informazioni su un oggetto che il sistema sa di non conoscere, ma l'impossibilità di formarsi persino l'ipotesi che certi stati interni esistano. Il "non sapere come si fa" non è reticenza o ineffabilità mistica: è la conseguenza precisa di $KC(v_expertise) > K$ dove K è il budget del prior metacognitivo.

La previsione verificabile: la correlazione tra la difficoltà di auto-descrizione dell'expertise in un dominio e la complessità algoritmica della struttura di quel dominio nello spazio di attivazione dovrebbe essere positiva e significativa. I domini dove l'expertise è facilmente verbalizzabile hanno $KC(v_expertise)$ bassa; i domini dove è difficile hanno KC alta. La difficoltà del "saper dire come" è una proprietà della complessità relativa tra struttura del substrato e capacità rappresentativa del meta-modello.

IX.4 Il problema del frame come struttura della metacognizione incrementale

Il *problema del frame* è stato formulato da McCarthy e Hayes (1969) nel contesto dell'AI simbolica: come si rappresenta ciò che non cambia in seguito a un'azione, senza enumerare esplicitamente tutto l'invariato?

Nel contesto metacognitivo emerge in forma più profonda. Quando il sistema aggiorna il prior π^0_t in risposta a un'osservazione o_t , produce π^0_{t+1} . L'aggiornamento bayesiano ottimale riguarda solo i parametri θ rilevanti per o_t ; i parametri irrilevanti dovrebbero restare invariati. Il *problema del frame metacognitivo* è: come il sistema determina quali aspetti di θ non sono toccati dall'aggiornamento, senza verificare esplicitamente ogni componente?

Formalmente si decompone $\Theta = \Theta_{\text{rilevante}}(o) \oplus \Theta_{\text{invariante}}(o)$. Il problema del frame chiede di determinare questa decomposizione per ogni osservazione o . Nel caso generale, ciò è computazionalmente equivalente al problema della fermata: determinare la struttura causale del proprio substrato non è decidibile algebricamente.

IX.5 Soluzioni approssimate e loro garanzie formali

Si identificano tre strategie, ciascuna con bound sull'errore di frame E_{frame} :

La *conservatività locale* assume rilevante solo il vicinato di raggio r nel grafo causale di Θ :

$$E_{\text{frame}}(r) \leq C_{\text{grafo}} \cdot (1 - \alpha)^r$$

con $\alpha \in (0,1)$ coefficiente di attenuazione e C_{grafo} costante strutturale. Per $r = O(\text{diametro del grafo})$ l'errore è trascurabile ma il costo è $O(n^2)$ nella dimensione del sistema.

La *persistenza per default* assume invarianza di tutto ciò non esplicitamente osservato:

$$E_{\text{frame}}(\text{default}) \leq \Sigma_{\{\theta \text{ non osservato}\}} |d\theta/dt| \cdot \Delta t$$

l'integrale delle variazioni non osservate nell'intervallo Δt . Il *change blindness* cognitivo è la sua manifestazione comportamentale.

La *modularità gerarchica* assume struttura modulare con influenze causali inter-modulo deboli:

$$E_{\text{frame}}(\text{moduli}) \leq C_{\text{inter}} \cdot \epsilon_{\text{inter}}$$

con ϵ_{inter} forza media delle connessioni inter-modulo. Per sistemi quasi-modulari l'errore è piccolo e il costo computazionale è lineare nel numero di moduli.

IX.6 Il problema del frame e il malinteso dinamico: connessione con la prima monografia

Il malinteso comunicativo della prima monografia — divergenza $D_{\text{KL}}(\pi^{\wedge}B_A \parallel \pi^{\wedge}B_{\text{reale}})$ tra il modello di A su B e la struttura effettiva di B — si arricchisce di una dimensione temporale attraverso il problema del frame.

Quando B compie l'azione a , A deve aggiornare il proprio modello di B. L'aggiornamento ottimale riguarda solo i parametri del modello di B rilevanti per a ; quelli irrilevanti dovrebbero restare invariati. Ma identificarli è soggetto al problema del frame. Il *malinteso dinamico* che ne emerge non è la divergenza tra stati iniziali (il caso classico della prima monografia) ma la divergenza tra traiettorie di aggiornamento:

$$d/dt D_{\text{KL}}(\pi^{\wedge}B_A(t) \parallel \pi^{\wedge}B_{\text{reale}}(t)) = D_{\text{frame}}(t) + D_{\text{comunicazione}}(t)$$

dove $D_{\text{frame}}(t) \geq 0$ è il contributo degli errori di frame e $D_{\text{comunicazione}}(t)$ è il contributo degli errori di codifica-decodifica della prima monografia. Il termine D_{frame} produce una deriva sistematica che cresce nel tempo anche quando i modelli iniziali sono accurati e la comunicazione è perfetta. Il punto fisso del gioco comunicativo della prima monografia acquista un termine di errore residuo $D_{\text{frame}} > 0$ irriducibile per via puramente comunicativa.

IX.7 La struttura diagonale come filo del ciclo: sintesi

Il ciclo di quattro monografie è un'esplorazione della struttura diagonale in quattro domini della cognizione.

Nella prima monografia — il malinteso — la diagonalizzazione emerge nella mutua modellazione: A modella B che modella A che modella B, con il malinteso come fallimento della convergenza al punto fisso del gioco.

Nella seconda monografia — il substrato subpropositivo — la diagonalizzazione emerge nella auto-attivazione della rete neurale: lo stato $v \in \mathbb{R}^n$ corrispondente al "pensare a se stessi" è il punto fisso della mappa di attivazione quando l'input è una rappresentazione del sistema stesso.

Nella terza monografia — la metacognizione — la diagonalizzazione emerge nei teoremi di impossibilità per l'auto-conoscenza e nella struttura del lemma diagonale: l'auto-rappresentazione completa produce invariabilmente il punto fisso impossibile. La geometria è quella dei frame GL.

Nella quarta monografia — l'attenzione — la diagonalizzazione emergerà nella auto-attenzione: il meccanismo che seleziona i prior da aggiornare è esso stesso soggetto all'aggiornamento che seleziona. Il punto fisso stabile dell'auto-attenzione è la struttura formale dell'identità cognitiva come attrattore; il punto fisso instabile è la fonte di distorsione sistematica dell'aggiornamento collettivo.

Il ciclo è unitario: il teorema di Lawvere è lo schema formale che unifica le quattro manifestazioni

Appendici tecniche

Appendice A — Il teorema di Lawvere nei topos: la versione categoriale

La versione più generale del teorema di punto fisso di Lawvere si enuncia nella categoria dei topos — la struttura categoriale che generalizza la teoria degli insiemi mantenendo sufficiente struttura logica per derivare i risultati di impossibilità.

Un topos è una categoria $\mathcal{T}$ con prodotti finiti, un classificatore dei sottoggetti Ω , e oggetti esponenziali B^A per ogni coppia di oggetti A, B . Il classificatore Ω è l'analogo dell'insieme $\{0,1\}$ dei valori di verità: ogni morfismo $A \rightarrow \Omega$ corrisponde a un "predicato" su A .

Teorema di Lawvere (versione categoriale): In un topos $\mathcal{T}$, se esiste un morfismo suriettivo $\varphi : A \twoheadrightarrow B^A$ (un epimorfismo dall'oggetto A allo spazio degli endomorfismi di A in B), allora ogni endomorfismo $f : B \rightarrow B$ ha un punto fisso.

La dimostrazione è identica alla versione insiemistica: si costruisce la diagonale $d = f \circ \text{eval} \circ (\varphi \times \text{id}_A) : A \rightarrow B$, dove $\text{eval} : B^A \times A \rightarrow B$ è il

morfismo di valutazione, e si usa la suriettività di ϕ per trovare a^* con $\phi(a^*) = d$.

Le istanze nei topos diversi dalla categoria degli insiemi producono risultati di impossibilità in contesti diversi. Nel topos dei fasci su uno spazio topologico — il contesto naturale per la logica modale e la semantica Kripke — il teorema di Lawvere produce risultati di incompletezza per sistemi modali, compresi i sistemi di belief revision che formalizzano la metacognizione nei modelli di epistemic logic. Nel topos dei presheaf su una categoria di modi di presentazione — il contesto della semantica dei tipi dipendenti — il teorema produce l'incompletezza dell'auto-tipizzazione, rilevante per i sistemi di tipi riflessivi nella teoria della programmazione.

Appendice B — La teoria bayesiana gerarchica: costruzione precisa del limite proiettivo

La costruzione precisa del limite proiettivo della torre bayesiana gerarchica richiede la specificazione della categoria in cui il limite viene preso. Si lavora nella categoria Meas delle misure di probabilità su spazi misurabili, con morfismi dati dagli operatori di marginalizzazione.

Gli oggetti della torre sono:

$$\begin{aligned}\Pi^0 &= \text{Prob}(\Theta) && \text{— misure di probabilità su } \Theta \\ \Pi^1 &= \text{Prob}(\Pi^0) && \text{— misure su misure su } \Theta \\ \Pi^k &= \text{Prob}(\Pi^{k-1}) && \text{— } k\text{-esimo livello}\end{aligned}$$

I morfismi di marginalizzazione sono:

$$\text{marg}_{\{k\}} : \Pi^k \rightarrow \Pi^{k-1}, \quad \text{marg}_{\{k\}}(\mu)(A) = \int_{\Pi^{k-1}} \pi(A) d\mu(\pi)$$

Questo sistema proiettivo $(\Pi^k, \text{marg}_{\{k\}})_{\{k \geq 0\}}$ ha un limite proiettivo $\Pi^\infty = \lim\{\leftarrow\} \Pi^k$ nella categoria Meas, la cui esistenza è garantita dal teorema di Kolmogorov per sistemi proiettivi coerenti di misure di probabilità su spazi polacchi. La condizione di coerenza è:

$$\text{marg}_{\{k\}} \circ \text{marg}_{\{k+1\}} = \text{marg}_{\{k-1\}} \quad \text{per ogni } k \geq 1$$

che è soddisfatta per costruzione dalla definizione di marginalizzazione iterata.

Il limite Π^∞ è uno spazio di misure di probabilità su $\Theta^\infty = \Theta \times \Pi^0 \times \Pi^1 \times \dots$, e un elemento di Π^∞ è una distribuzione consistente su tutti i livelli della torre: la struttura formale del "conoscitore bayesiano ideale a tutti i livelli" di cui la metacognizione finita è una troncatura approssimata.

Proposizione: La profondità metacognitiva ottimale K^* per un sistema con budget computazionale B_{meta} è:

$$K^* = \operatorname{argmax}_{\{K\}} [R_{\text{meta}}(D_K) - \lambda \cdot C(K)]$$

dove D_K è la distorsione della torre troncata al livello K , $R_{\text{meta}}(D_K)$ è la riduzione della distorsione dal livello K rispetto al livello $K-1$, $C(K)$ è il costo computazionale del mantenimento del livello K , e λ è il costo marginale dell'elaborazione metacognitiva. La profondità ottimale bilancia il guadagno in accuratezza metacognitiva con il costo di mantenimento del livello aggiuntivo.

Appendice C — Il risultato di no-free-lunch per la metacognizione: derivazione

Il teorema di no-free-lunch di Wolpert e Macready (1997) nella versione pertinente per la metacognizione si enuncia come segue.

Sia $\mathbb{F}$ l'insieme di tutti i possibili problemi cognitivi — tutte le funzioni da stimoli a risposte ottimali su un dominio finito X — e sia $\mathcal{M}$ l'insieme di tutti i meta-modelli metacognitivi — tutte le distribuzioni π^0 sullo spazio Θ dei parametri cognitivi. Per un meta-modello $m \in \mathcal{M}$ e un problema $f \in \mathbb{F}$, sia $P(f, m, t)$ la probabilità di produrre una risposta corretta al problema f con il meta-modello m dopo t passi di allocazione.

Teorema (No-Free-Lunch per la metacognizione): Per qualsiasi coppia di meta-modelli $m_1, m_2 \in \mathcal{M}$ e per qualsiasi $t > 0$:

$$\sum_{\{f \in \mathbb{F}\}} P(f, m_1, t) = \sum_{\{f \in \mathbb{F}\}} P(f, m_2, t)$$

mediando uniformemente su tutti i problemi cognitivi, tutti i meta-modelli hanno la stessa performance attesa.

Derivazione: Si sfrutta la simmetria della misura uniforme su $\mathbb{F}$. Fissati m_1 e t , si costruisce una biiezione $\sigma_{\{m_1, m_2\}} : \mathbb{F} \rightarrow \mathbb{F}$ tale che $P(f, m_1, t) = P(\sigma(f), m_2, t)$ per ogni f . Tale biiezione esiste perché l'insieme dei problemi cognitivi è chiuso per permutazione delle etichette di X , e il meta-modello opera sulle strutture di X in modo che può essere riformulato come un problema su X permutato. La biiezione conserva la misura uniforme, quindi la somma è conservata. $\square$

Il corollario per la pratica metacognitiva è che la scelta del meta-modello ottimale richiede un prior sulla distribuzione dei problemi cognitivi che il sistema incontrerà: senza questa informazione a priori, nessun meta-modello è preferibile a un altro in attesa.

Appendice D — Il problema del frame: formalizzazione e complessità

D.1 Formalizzazione nel calcolo delle situazioni

Nel calcolo delle situazioni (McCarthy & Hayes, 1969), un'azione A trasforma la situazione S in $\text{Result}(A, S)$. I *fluent* F sono proprietà che possono variare tra situazioni. Il problema del frame chiede di caratterizzare algebricamente l'insieme $\{F : F \text{ invariante sotto } A\}$ per ogni azione A .

La soluzione mediante *assiomi di frame* — "se F è vero in S e A non modifica F , allora F è vero in $\text{Result}(A, S)$ " — richiede determinare se A modifica F . Nella sua forma generale, questo è Π_2^P -completo nella gerarchia polinomiale.

Per il *frame metacognitivo*: determinare l'insieme $\{\theta \in \Theta : \theta \text{ invariante sotto osservazione } o\}$ è, nel caso generale senza ipotesi strutturali su Θ , computazionalmente equivalente al problema della fermata. Con ipotesi di sparsità sul grafo causale di Θ , il problema scende a polinomiale nella dimensione del sistema.

D.2 Analisi della complessità delle strategie approssimate

La conservatività locale con raggio r ha complessità $O(n \cdot r^d)$ dove d è la dimensione del grafo causale e n il numero di parametri. Per $d = 2$ (struttura planare) e $r = O(\log n)$ l'errore decade polinomialmente con n a costo $O(n \log^2 n)$.

La persistenza per default ha complessità $O(k_{\text{obs}})$ dove k_{obs} è il numero di parametri direttamente osservati. È computazionalmente ottimale ma massimizza l'errore di frame per sistemi in rapida evoluzione.

La modularità gerarchica ha complessità $O(n_{\text{moduli}} \cdot k_{\text{modulo}})$ dove n_{moduli} è il numero di moduli e k_{modulo} la dimensione media. Per strutture gerarchiche bilanciate questo è $O(n \log n)$.

D.3 Connessione con il teorema di Rice

Il teorema di Rice (Capitolo II) afferma che nessuna proprietà semantica non banale dei programmi è decidibile. Il problema del frame metacognitivo è un'istanza di questo risultato quando si chiede: "il parametro θ ha la proprietà di essere invariante sotto l'osservazione o ?" — questa è una proprietà semantica di θ inteso come programma, e per il teorema di Rice non è decidibile algebricamente nel caso generale.

La modularità cognitiva è la soluzione strutturale al problema del frame che elude il teorema di Rice: restringendo la classe delle proprietà considerate (da "semantiche arbitrarie" a "confinamento a moduli specifici"), si rende il problema trattabile a scapito della generalità.

Appendice E — Logica GL: derivazioni fondamentali

E.1 Verifica degli assiomi nei frame GL

Si verifica che gli assiomi (K), (4), (W) siano validi nella classe dei frame transitivi noetheriani.

(K): $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$. Sia $w \models \Box(\varphi \rightarrow \psi)$ e $w \models \Box\varphi$. Per ogni v con wRv : $v \models \varphi \rightarrow \psi$ e $v \models \varphi$, quindi $v \models \psi$. Dunque $w \models \Box\psi$. Valido in tutti i frame.

(4): $\Box\varphi \rightarrow \Box\Box\varphi$. Sia $w \models \Box\varphi$. Per ogni v con wRv : $v \models \varphi$. Si vuole mostrare che per ogni v con wRv , $v \models \Box\varphi$. Sia v con wRv e sia u con vRu : per transitività wRu , quindi $u \models \varphi$. Dunque $v \models \Box\varphi$. Valido nei frame transitivi.

(W): $\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$. Sia $w \models \Box(\Box\varphi \rightarrow \varphi)$. Si vuole $w \models \Box\varphi$, ovvero che per ogni v con wRv , $v \models \varphi$. Si considera l'insieme $S = \{v : wRv \text{ e } v \not\models \varphi\}$ e si suppone S non vuoto. Per noetherianità, S ha un elemento R -massimale m : nessun $u \in S$ è tale che mRu . Per costruzione $m \not\models \varphi$. Ma da $w \models \Box(\Box\varphi \rightarrow \varphi)$ e wRm si ha $m \models \Box\varphi \rightarrow \varphi$. Poiché $m \not\models \varphi$, si conclude $m \not\models \Box\varphi$: esiste u con mRu e $u \not\models \varphi$. Ma allora wRm e mRu implica wRu per transitività, quindi $u \in S$. Poiché mRu con $u \in S$, m non è R -massimale in S : contraddizione. Quindi S è vuoto e $w \models \Box\varphi$. Valido nei frame transitivi noetheriani.

E.2 Il frame canonico di GL e la completezza di Segerberg

Il frame canonico $C(GL)$ ha come mondi gli insiemi massimali consistenti di GL (teorie complete). La relazione di accessibilità è:

$$w R_c v \iff \{\varphi : \Box\varphi \in w\} \subseteq v$$

Transitività: siano $w R_c v$ e $v R_c u$. Sia $\Box\varphi \in w$. Da $w R_c v$, $\varphi \in v$; da (4) in GL, $\Box\varphi \rightarrow \Box\Box\varphi$, quindi $\Box\Box\varphi \in w$, e poiché $w R_c v$ si ha $\Box\varphi \in v$; da $v R_c u$ si ha $\varphi \in u$. Dunque $w R_c u$.

Noetherianità: si suppone esista una catena infinita $w_0 R_c w_1 R_c w_2 R_c \dots$. Si costruisce φ tale che $\Box\varphi \notin w_0$ ma $\varphi \in v$ per ogni v con $w_0 R_c v$ — contraddicendo l'assioma (W). La costruzione usa la formula diagonale applicata alla catena infinita, che produce una formula "vero ad ogni passo ma non uniformemente dimostrabile" — esattamente il tipo di formula che (W) esclude.

E.3 Derivazione del teorema di Löb dal lemma diagonale

La derivazione presentata nel Capitolo VIII (paragrafo VIII.6) usa il punto fisso $\ell \leftrightarrow (\Box\ell \rightarrow \varphi)$. Si verifica in dettaglio che $PA \vdash \Box\ell \rightarrow \Box\varphi$.

Da $\ell \leftrightarrow (\Box\ell \rightarrow \varphi)$ per il lemma diagonale, PA dimostra questa equivalenza. Applicando la necessitazione e (K):

$$PA \vdash \Box(\ell \leftrightarrow (\Box\ell \rightarrow \varphi))$$

$$PA \vdash \Box\ell \leftrightarrow \Box(\Box\ell \rightarrow \varphi) \quad (\text{da (K) applicato all'equivalenza})$$

$$PA \vdash \Box(\Box\ell \rightarrow \varphi) \leftrightarrow (\Box\Box\ell \rightarrow \Box\varphi) \quad (\text{da (K)})$$

Da (4): $PA \vdash \Box\ell \rightarrow \Box\Box\ell$. Componendo:

$$PA \vdash \Box\ell \rightarrow \Box\Box\ell \rightarrow \Box\varphi$$

Dall'ipotesi $PA \vdash \Box\varphi \rightarrow \varphi$ e da $\Box\varphi$ appena ottenuto: $PA \vdash \varphi$. $\square$

Appendice F — Tavola degli ordinali proof-teorici e struttura degli epsilon-numeri

F.1 Costruzione degli epsilon-numeri in forma normale di Cantor

Ogni ordinale $\alpha < \varepsilon_0$ ha una rappresentazione unica in *forma normale di Cantor in base ω* :

$$\alpha = \omega^{\alpha_1} \cdot k_1 + \omega^{\alpha_2} \cdot k_2 + \dots + \omega^{\alpha_n} \cdot k_n$$

con $\alpha > \alpha_1 > \alpha_2 > \dots > \alpha_n \geq 0$ e $k_i \in \mathbb{N}_{>0}$. La chiusura per moltiplicazione da sinistra — ω^α — non termina per $\alpha < \varepsilon_0$: ε_0 è il primo ordinale che è il proprio esponente, $\varepsilon_0 = \omega^{\varepsilon_0}$.

La sequenza degli epsilon-numeri è costruita iterando la funzione φ_1 della gerarchia di Veblen:

$$\varepsilon_0 = \varphi_1(0), \quad \varepsilon_1 = \varphi_1(1), \quad \varepsilon_\alpha = \varphi_1(\alpha)$$

La funzione $\varphi_0(\alpha) = \omega^\alpha$ enumera i punti fissi di sé stessa; $\varphi_1(\alpha)$ enumera i punti fissi di φ_0 ; $\varphi_2(\alpha)$ enumera i punti fissi di φ_1 ; e così via. Il limite della gerarchia $\varphi_\alpha(0)$ per $\alpha \rightarrow \omega^2$ è l'ordinale di Feferman-Schütte Γ_0 .

F.2 Tavola degli ordinali proof-teorici

Sistema	Ordinale proof-teorico
---------	------------------------

Q (aritmetica di Robinson)	ω
PRA (ricorsiva primitiva)	ω^ω
IS_1 (induzione Σ_1)	ω^ω
IS_n (induzione Σ_n , $n>1$)	ω^{ω^n}
PA (Peano completa)	ε_0
PA + Con(PA)	ε_1
PA + Con(PA + Con(PA))	ε_2
ACA_0 (aritmetica secondo ordine base)	ε_0
ACA (aritmetica secondo ordine + induz.)	ε_ω
ATR_0 (ricorsione transfinita)	Γ_0
ATR (ricorsione transfinita + induz.)	$\Gamma_0 \cdot \omega$
$\Pi^1_1\text{-}CA_0$	$\psi_0(\Omega^\Omega)$
$\Pi^1_1\text{-}CA$	$\psi_0(\Omega^{\Omega \cdot \omega})$
ZFC (lower bound stimato)	$\gg \psi_0(\Omega_\omega)$

F.3 Corrispondenza con la torre metacognitiva

Livello bayesiano	Capacità metacognitiva	Ordinale
π^0 (K=0)	modello di base dei propri parametri	ε_0
π^1 (K=1)	certifica coerenza del proprio K=0	ε_1
π^2 (K=2)	certifica coerenza del proprio K=1	ε_2
π^k (K=k)	certifica coerenza del proprio K=k-1	ε_k
Π^∞ (K= ∞ , ideale)	auto-conoscenza completa (infattibile)	$\zeta_0 = \varepsilon_{\{\varepsilon_{\{\dots\}}\}}$

Dove $\zeta_0 = \varphi_2(0)$ è il primo punto fisso di $\alpha \mapsto \varepsilon_\alpha$ — l'ordinale della torre metacognitiva ideale infinita, raggiungibile da qualsiasi sistema computabile ma formalmente ben definito come limite della gerarchia.

Bibliografia

Le opere qui elencate sono quelle su cui il quadro formale della monografia si fonda direttamente, con inclusione selettiva dei risultati sui teoremi di impossibilità che il ciclo di saggi presuppone già trattati.

Fondamenti bayesiani gerarchici e metacognizione

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1), 1-68. Il testo fondativo del probabilismo soggettivo bayesiano — la struttura di credenza come distribuzione di probabilità soggettiva — che è la base del Capitolo I. La rappresentazione gerarchica delle credenze su credenze è l'estensione naturale del framework di de Finetti.

Jaynes, E.T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge. Il testo che sviluppa il programma bayesiano come teoria normativa dell'inferenza razionale. Il Capitolo 18 sulle distribuzioni di secondo ordine e il Capitolo 19 sull'assegnazione dei prior sono direttamente rilevanti per la costruzione del meta-prior del Capitolo I.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). *Bayesian Data Analysis*, terza edizione. CRC Press, Boca Raton. Il testo di riferimento per la pratica dell'inferenza bayesiana gerarchica — il framework computazionale del Capitolo I della monografia.

Teoremi di impossibilità e struttura logica

Lawvere, F.W. (1969). Diagonal arguments and cartesian closed categories. *Category Theory, Homology Theory and Their Applications II*, Lecture Notes in Mathematics, 92, 134-145. L'articolo originale che unifica i risultati di incompletezza attraverso il teorema del punto fisso nella categoria degli insiemi e nei topos chiusi cartesiani. Il nucleo teorico del Capitolo II della monografia.

Yanofsky, N.S. (2003). A universal approach to self-referential paradoxes, incompleteness and fixed points. *Bulletin of Symbolic Logic*, 9(3), 362-386. Un'esposizione accessibile e sistematica del programma di unificazione di Lawvere, che mostra esplicitamente come i teoremi di Cantor, Russell, Gödel, Tarski, Halting Problem e molti altri siano istanze dello stesso schema. Testo di riferimento per la tassonomia del Capitolo II.

Löb, M.H. (1955). Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20(2), 115-118. L'articolo originale che dimostra il teorema che porta il nome di Löb

— il limite dell'auto-fiducia epistemica. La derivazione originale è elegante e occupa meno di quattro pagine.

Rice, H.G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, 74(2), 358-366. L'articolo originale del teorema di Rice sull'indcidibilità delle proprietà semantiche non banali dei programmi.

Chaitin, G.J. (1975). A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3), 329-340. Il testo che stabilisce i bound sulla complessità di Kolmogorov auto-dimostrabile — la versione quantitativa del risultato di Gödel nella formulazione del Capitolo II.

Calibrazione metacognitiva e sue deviazioni

Kruger, J., Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. L'articolo che documenta l'effetto Dunning-Kruger — la miscalibrazione metacognitiva sistematica nelle direzioni descritte nel Capitolo I.

Fleming, S.M., Dolan, R.J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B*, 367(1594), 1338-1349. La revisione delle basi neurali della metacognizione — i substrati biologici del meta-modello π^0 nelle aree prefrontali — che connette il framework formale del Capitolo I con la neurobiologia.

Metcalf, J., Shimamura, A.P. (Eds.) (1994). *Metacognition: Knowing about Knowing*. MIT Press, Cambridge MA. Il volume che raccoglie i contributi fondamentali alla psicologia della metacognizione — la base empirica del framework teorico della monografia.

Teoria della compressione e codifica rate-distortion

Berger, T. (1971). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs NJ. Il testo fondativo della teoria del tasso-distorsione — il framework quantitativo del Capitolo III della monografia. Il Capitolo 4 sulla codifica delle sorgenti con distorsione e il Capitolo 6 sulla soluzione water-filling sono i riferimenti diretti.

Blahut, R.E. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4), 460-473. L'algoritmo di Blahut-Arimoto per il calcolo numerico della funzione di tasso-distorsione — la base computazionale delle stime nel Capitolo III.

Teoria dell'allocazione delle risorse cognitive

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624-652. Lo studio del ruolo della corteccia anteriore cingolata nel monitoraggio del conflitto cognitivo e

nell'allocazione delle risorse di controllo — la base neurobiologica del meccanismo formalizzato nel Capitolo IV.

Shallice, T., Cooper, R.P. (2011). *The Organisation of Mind*. Oxford University Press, Oxford. Il testo che elabora la teoria dei Supervisory Attentional Systems — la struttura di controllo cognitivo di livello superiore che corrisponde, nel framework formale, al meta-modello come struttura di controllo della compressione.

Impossibilità collettiva: teoremi di Arrow e Sen

Arrow, K.J. (1951). *Social Choice and Individual Values*. Wiley, New York. Il testo originale che dimostra l'impossibilità dell'aggregazione delle preferenze individuali in una preferenza collettiva coerente soddisfacente le quattro condizioni — il risultato di impossibilità collettiva il cui schema è connesso ai risultati individuali nel Capitolo V.

Sen, A.K. (1970). The impossibility of a Paretian liberal. *Journal of Political Economy*, 78(1), 152-157. Il risultato di Sen sull'impossibilità di una funzione di scelta sociale che soddisfi simultaneamente il principio di Pareto e i diritti individuali minimi — l'altra istanza dei teoremi di impossibilità per l'aggregazione collettiva.

Genest, C., Zidek, J.V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1), 114-135. La revisione sistematica dei risultati di impossibilità per l'aggregazione bayesiana delle distribuzioni di credenza — il risultato che estende il teorema di Arrow al contesto delle distribuzioni di probabilità nel Capitolo V.

Il problema dell'opinione del comitato bayesiano

Hylland, A., Zeckhauser, R. (1979). The impossibility of Bayesian group decision making with separate aggregation of beliefs and values. *Econometrica*, 47(6), 1421-1436. Il risultato di impossibilità per l'aggregazione bayesiana coerente delle credenze di gruppo — la versione bayesiana del teorema di Arrow trattata nel Capitolo V.

No-free-lunch e ottimizzazione

Wolpert, D.H., Macready, W.G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67-82. Il testo originale del teorema di no-free-lunch per l'ottimizzazione — il risultato il cui schema è applicato alla metacognizione nella Appendice C e nel Capitolo II.

Sistemi cognitivi collettivi e cascate informazionali

Bikhchandani, S., Hirshleifer, D., Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992-1026. Il modello formale delle cascate informazionali — il meccanismo attraverso cui agenti razionali convergono a credenze sub-ottimali seguendo l'osservazione del comportamento altrui — trattato nel Capitolo V.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford University Press, Oxford. Il testo che sviluppa l'epistemologia sociale come disciplina autonoma — il framework filosofico per la metacognizione collettiva del Capitolo V.

Logica GL e teoria della dimostrabilità

Boolos, G. (1993). *The Logic of Provability*. Cambridge University Press, Cambridge. Il testo di riferimento per la logica GL: semantica di Kripke, teorema di Solovay con dimostrazione completa, applicazioni ai fondamenti.

Boolos, G. (1979). *The Unprovability of Consistency: An Essay in Modal Logic*. Cambridge University Press. La prima monografia sistematica sulla connessione tra il secondo teorema di incompletezza e GL.

Solovay, R.M. (1976). Provability interpretations of modal logic. *Israel Journal of Mathematics*, 25(3-4), 287-304. L'articolo originale della completezza aritmetica di GL — il teorema centrale del Capitolo VI.

Segerberg, K. (1971). *An Essay in Classical Modal Logic*, 3 voll. Filosofiska Studier 13, Uppsala. La completezza semantica di K4 e delle logiche modali connesse.

Lindström, P. (1997). *Aspects of Incompleteness*. Springer, Berlino. Trattazione tecnica di logica della dimostrabilità, incompletezza e logica modale.

Smoryński, C. (1985). *Self-Reference and Modal Logic*. Springer, New York. Applicazioni sistematiche del lemma diagonale alla logica modale: derivazioni complete per Gödel, Tarski e Löb come casi particolari.

Analisi proof-teorica e ordinali

Gentzen, G. (1936). Die Widerspruchsfreiheit der reinen Zahlentheorie. *Mathematische Annalen*, 112, 493-565. La dimostrazione originale di $\text{Con}(\text{PA})$ mediante $\text{TI}(\varepsilon_0)$.

Pohlers, W. (2009). *Proof Theory: The First Step into Impredicativity*. Springer, Berlino. Derivazione degli ordinali proof-teorici per i sistemi dell'aritmetica di secondo ordine inclusi ATR_0 e $\Pi^1_1\text{-CA}_0$.

Rathjen, M. (2006). The art of ordinal analysis. *International Congress of Mathematicians*, Vol. II, 45-69, EMS, Zürich. Panoramica dell'analisi ordinale moderna da PRA ai sistemi impredicativi oltre Γ_0 .

Feferman, S. (1964). Systems of predicative analysis. *Journal of Symbolic Logic*, 29(1), 1-30. Il programma predicativista con l'identificazione di Γ_0 come limite della matematica predicativa — la barriera metacognitiva del Capitolo VII.

Schütte, K. (1965). Eine Grenze für die Beweisbarkeit der transfiniten Induktion in der verzweigten Typenlogik. *Archiv für mathematische Logik und Grundlagenforschung*, 7(1-2), 45-60. Il risultato indipendente di Schütte che identifica Γ_0 come ordinale di Feferman-Schütte.

Il lemma diagonale

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38(1), 173-198. L'articolo originale — fonte primaria del lemma diagonale nella sua prima incarnazione.

Carnap, R. (1934). *Logische Syntax der Sprache*. Springer, Wien. Formulazione indipendente del lemma diagonale da parte di Carnap.

Prior di Solomonoff e complessità algoritmica

Solomonoff, R.J. (1964). A formal theory of inductive inference. *Information and Control*, 7(1), 1-22 e 7(2), 224-254. Gli articoli fondativi del prior universale.

Li, M., Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*, terza edizione. Springer, New York. Il testo di riferimento per KC, il prior di Solomonoff, e i legami con Gödel e Chaitin. Capitolo 4 (prior universale) e Capitolo 7 (connessioni con incompletezza).

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, Berlino. Il sistema AIXI che usa il prior di Solomonoff — modello teorico di agente razionale ideale computazionalmente infattibile per le ragioni del Capitolo IX.

Il problema del frame

McCarthy, J., Hayes, P.J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer, B., Michie, D. (Eds.), *Machine Intelligence*, vol. 4, 463-502, Edinburgh University Press. L'articolo originale del problema del frame nel calcolo delle situazioni.

Shanahan, M. (1997). *Solving the Frame Problem*. MIT Press, Cambridge MA. Monografia sistematica sulle soluzioni approssimate al problema del frame — base delle strategie formalizzate nell'Appendice D.

Lifschitz, V. (2015). The frame problem, then and now. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, AAAI Press, 4327-4331. Revisione dello stato dell'arte con connessioni alla pianificazione e al ragionamento non-monotono.

Connessioni con la prima e seconda monografia

Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y. (1995). *Reasoning About Knowledge*. MIT Press, Cambridge MA. Il testo fondativo della logica epistemica formale — framework che connette la semantica di Kripke con il ragionamento sugli stati mentali degli agenti della prima monografia.

Bikhchandani, S., Hirshleifer, D., Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5),

992-1026. Le cascate informazionali come meccanismo di convergenza sub-ottimale del meta-modello collettivo del Capitolo V.